

A THEORETICALLY LOSSLESS METHOD FOR DERIVING EMPIRICAL ORTHOGONAL
FUNCTIONS FROM UNEVENLY SAMPLED DATA

A Thesis

by

CHRISTOPHER M. DUPUIS

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,
Committee Members,

Head of Department,

Courtney Schumacher
Robert Korty
Steven DiMarco
Ping Yang

May 2016

Major Subject: Atmospheric Sciences

Copyright 2016 Christopher M. Dupuis

ABSTRACT

The Lomb-Scargle discrete Fourier transform (LSDFT) is a fairly popular technique for analyzing time series within the astronomy community. However, this algorithm is largely unknown in many other disciplines, despite many potential applications. In particular, the atmospheric sciences stand to benefit substantially from implementing it, since much of the corpus of observational data is irregularly sampled. In this study, a solution for empirical orthogonal functions (EOFs) based on irregularly sampled data is derived from the LSDFT. It is demonstrated that this particular algorithm has no hard limit on its accuracy, and yields results comparable to those of complex Hilbert EOF analysis. Three LSDFT algorithms are compared in terms of their performance in evaluating EOFs for data from the Tropical Rainfall Measuring Mission. All three are shown to be able to capture the pattern of the diurnal cycle, and also show other consistent features in the zero to twelve day frequency band. The feasibility of implementing these algorithms is also investigated, and it is found that the programming language R is only about 2.2 ± 0.1 times as slow as CUDA C/C++ in this particular application.

TABLE OF CONTENTS

	Page
ABSTRACT.....	ii
TABLE OF CONTENTS.....	iii
LIST OF FIGURES.....	iv
LIST OF TABLES	v
1. INTRODUCTION	1
1.1 The Empirical Orthogonal Function Problem.....	4
1.2 Exact Nonequispaced Time Series Methods.....	18
1.3 Approximate Nonequispaced Time Series Methods	26
1.4 The Statistics of Rainfall.....	27
1.5 Computational Aspects.....	28
2. METHODS.....	30
2.1 Empirical Orthogonal Functions	30
2.2 The Lomb-Scargle Discrete Fourier Transform	32
2.3 Statistical Distribution of Rainfall.....	39
2.4 Computation and Benchmarking	44
3. RESULTS.....	46
3.1 Computational Results	72
4. DISCUSSION	73
5. CONCLUSIONS	77
REFERENCES	79

LIST OF FIGURES

	Page
Figure 1 Covariance is generally impossible to calculate directly from gappy time series data, but several alternatives exist.....	3
Figure 2 Workflows for calculating covariance using (left) the univariate integral and (right) the bivariate integral.....	31
Figure 3 The general shape of a rainfall cumulative distribution, not including the autoregressive term.....	43
Figure 4 WOSA results with theoretical AR1 distribution as from SS02 in blue, and the two-sigma confidence limits as dashed lines.....	47
Figure 5 The classic LSDFT analysis with a complex formulation over South America.	48
Figure 6 The WOSA method applied to the LSDFT (top) results in stronger relative amplitudes in most places, but appears to granulate spatial patterns in general, and (bottom) requires circular averaging of phases, which appears to granulate spatial patterns in phase as well.....	49
Figure 7 The WOSA method corrected for AR1 red noise estimates as per SS02.	50
Figure 8 EOF results from classic LSDFT analysis, over the zero to twelve day bandwidth, showing relative amplitude and phase.....	52
Figure 9 EOF results from WOSA analysis, showing relative amplitude and phase of the first four EOFs.....	57
Figure 10 Selected EOFs resulting from WOSA analysis with AR1 reduction.	61
Figure 11 The WOSA EOF analysis of the diurnal cycle.	64
Figure 12 An EOF analysis of just the semi-diurnal component of WOSA analysis results in a pattern consistent with the second EOF of the integrated analysis from Figure 9.....	65
Figure 13 EOF results from WOSA analysis on the TRMM data with a large, artificial gap.....	68

LIST OF TABLES

	Page
Table 1	Variances and Error Estimates for LSDFT EOF analysis 66
Table 2	Variances and Error Estimates for WOSA EOF analysis..... 66
Table 3	Variances and Error Estimates for AR1-reduced EOF analysis 67
Table 4	Timing of R and CUDA LSDFT algorithms..... 72

1. INTRODUCTION

Empirical orthogonal functions (or EOFs) are an important part of spatial analysis of oscillations, and are an especially pervasive technique in atmospheric science since the 1950s or so (North et al., 1982). However, they have largely been applied to model data sets, since there is currently no theoretically exact way to connect EOF analysis to nonequispaced data. “Nonequispaced” (alternatively “gappy,” “irregularly spaced,” or “unevenly spaced”) in this context refers to data that is sampled at irregular time intervals, a situation which typically arises for two reasons: 1) the sampling technique does not generally result in evenly spaced data, or 2) the sampling technique does not always lead to valid results, even when sampling is otherwise equispaced. An example of the first scenario would be paleoclimate records; since layering rates are variable through the record for many types of paleoclimate proxies, different periods will have corresponding variations in temporal resolution. An example of the second scenario would be telescope or satellite observations. Reasons for uneven spacing include the diurnal and annual cycles (which, in astronomy, can block the star of interest during daytime), the weather, telescope scheduling, and technical difficulties such as cosmic ray interference. In these cases, the reason for uneven spacing is a property extrinsic to the technique in question.

The fundamental difficulty with connecting unevenly spaced time series to the EOF problem lies in calculating the covariance matrix, each element of which is simply the covariance between the two time series it represents. EOFs are the eigenvectors of a covariance matrix, and the eigenvalues that result from the same solution represent the relative variances of the eigenvectors. Approximate methods do exist; however, they rely

on various theoretical frameworks that are most appropriate only for specific cases, usually equispaced data (Figure 1). The simplest of these methods is to use pairwise-complete observations. For some data sets (like in the second scenario described previously), the observation times for each pair of time series may be essentially the same. In this situation, covariance can be calculated directly after removing the incomplete observation pairs. There are also numerous interpolation methods, which are designed to fit unevenly spaced data into an even spacing. These both inherently result in data loss, for different reasons. Pairwise-complete observations by definition won't include every data point if there is a gap, and this problem can result in large data losses when sampling times between time series are significantly different. With interpolation, known errors are introduced, and interpolation to even spacing is particularly ill-suited to time series with large gaps, as are common in astronomy and atmospheric science. There is also the possibility that the variable in question follows an unusual statistical distribution, in which case, accurate interpolation may become much more complicated.

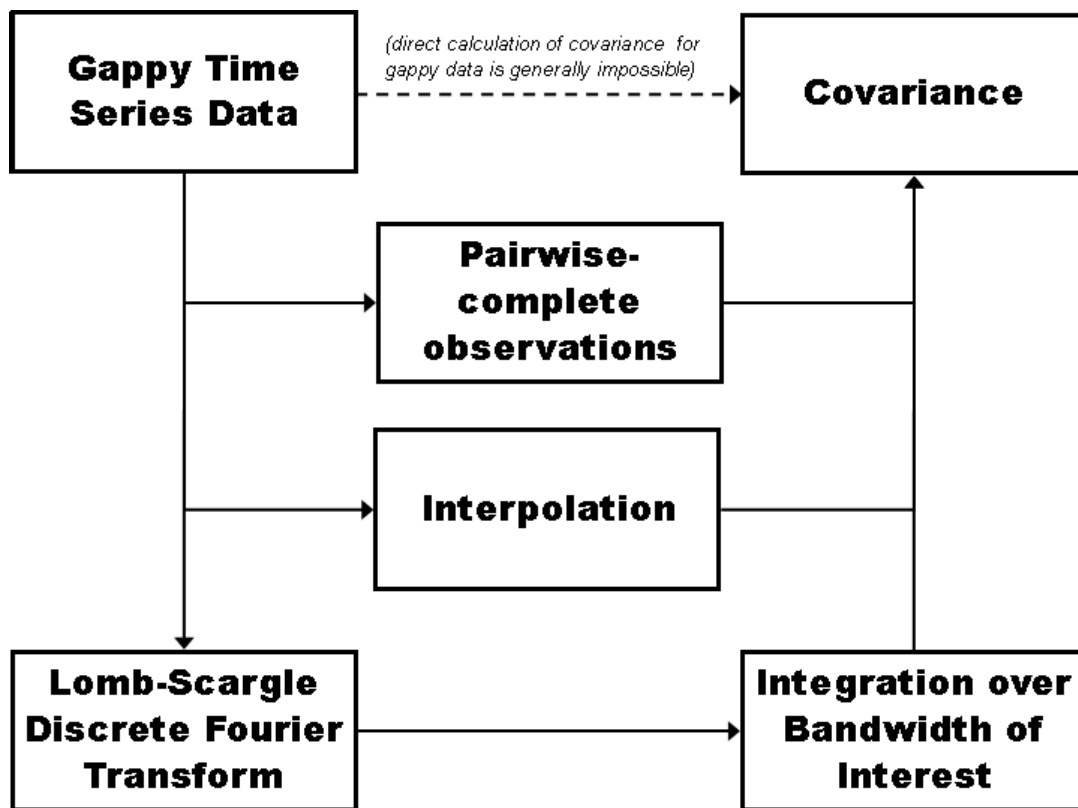


Figure 1: Covariance is generally impossible to calculate directly from gappy time series data, but several alternatives exist.

An additional pathway remains: to use spectral analysis. Spectral analysis allows for the use of all available data, and since there is no hard limit on how accurately a spectrum can be calculated, spectral analysis can be theoretically errorless in the best-case scenario. Spectral analysis can also manage large gaps in sampling times.

Time series spectral analysis has long been an integral component of observational and statistical science. Of particular note is the use of the Fourier transform, an integral transform that converts a function from the time domain to the frequency domain, wherein

periodicities can easily be observed. For a time series S as a function of time (t) and its Fourier transform \tilde{S} as a function of angular frequency ω are related by

$$\tilde{S}(\omega) = \int_0^\infty S(t) e^{-i\omega(t+t_{lag})} dt \approx \sum_{j=1}^{N_0} S(t_j) e^{-i\omega(t+t_{lag})}, \quad (1.1)$$

where t_{lag} is the autocorrelation lag time. Fourier transforms are typically normalized, often by a factor of $1/\sqrt{N}$ where N denotes the number of (valid) observations. Although lagged correlations are not included in this study, it is simple to see how the $e^{-i\omega t_{lag}}$ term propagates through the later equations. For analysis of data, a discretized Fourier transform (DFT) is usually used. Optimization of DFT algorithms has resulted in a family of routines known as Fast Fourier Transforms (FFTs), which reduce computation time from $O(N^2)$ to $O(N \ln(N))$, which are among the most popular scientific algorithms used today.

1.1 The Empirical Orthogonal Function Problem

Computationally, EOF analysis can be performed from at least three basic perspectives. The first perspective is to evaluate the EOFs as a stand-alone phenomenon. This point of view is probably the simplest to approach EOFs from, but it yields no information about how the EOFs may vary in time. The second perspective is to perform both EOF analysis and principle component analysis (PCA) as solutions to separate covariance matrix eigenvalue problems. The third perspective is to view EOFs, PCs, and the variance values as corresponding to the three matrices that result from singular value decomposition. The latter two of these perspectives include principle component analysis, so they can account for temporal variation in EOF strength. Choosing between these three points of view is mostly a matter of computational resources and research needs.

Additionally, EOFs can be calculated through two basic algorithms. The first is the classical algorithm, in which all the EOF values and vectors are solved simultaneously. Alternatively, the second algorithm calls for solving for the single largest EOF. This can be extended by iteratively solving for the largest EOF, subtracting it from the data, and reanalyzing the covariance matrix. This iterative method can be more efficient when only a few EOFs are desired from a large data set, particularly when considering that most EOFs correspond to negligible variances, and are therefore statistically meaningless and scientifically uninteresting.

Estimating the sampling errors of particular EOF variances is made possible by North's Rule of Thumb (North et al., 1982). An important consequence of these estimated errors is that EOFs with similar variances will be statistically unstable, and even if they collectively represent unrelated EOFs, their proximity may result in artificial degeneracies, which entails a mixing of spatial patterns. North et al. (1982) posits that EOF variances that differ by at least one standard error, as defined therein, can safely be considered distinct from each other. Since the error resulting from North's Rule of Thumb is a function of the number of data entries at a particular point, it takes a much larger amount of data to resolve two interacting EOFs with nearly the same variance than it does to resolve interacting EOFs that are somewhat more differentiated in terms of variance.

Among other quantities, EOFs can be related to normal modes of oscillations (North, 1984). North (1984) reviews several limitations of EOF analysis previously noted in other works, such as the fact that EOF spatial patterns are significantly affected by the spatial domain chosen, as well as the statistical instability of spatial patterns with similar variances (as noted in North et al., 1982). North (1984) notes that ideally, the field quantity in question has no frequency dependency, but this is not necessarily true for quantities like

vorticity. This provides motivation for defining a “four-dimensional” EOF notation. North (1984) begins by applying the Fourier transform to the scalar field of interest, and then notes that since the cross-spectrum between two spatial points is a Hermitian operator, it can be used as the kernel in an eigenvalue problem. It is shown that the form of such a covariance matrix is essentially the same as for the standard EOF problem, with the exception that the frequency-dependent EOFs are complex-valued. North (1984) further notes that a particular class of linear, stochastically forced systems with homogeneous boundary conditions includes several physical problems which result in exactly the kind of frequency-dependent EOFs described therein. These problems include the wave equation and the diffusion equation when they include a stochastic forcing function, as well as the barotropic vorticity equation for the stream function. However, the problems that can be solved with frequency-dependent EOFs are not limited to these cases. As North (1984) notes, the important thing is that the forcing variance is uniform across the field. This language assumes a normal distribution, but the basic requirement that the second moment (and possibly higher-order moments) is likely extensible to other statistical distributions. Another subset of these problems includes those described above, with random initial conditions and zero forcing instead of a stochastic forcing function. The evolution of Rossby waves is noted as one example of this class of problem.

North (1984) goes on to discuss the nature of the complex components, demonstrating that systems that are not purely Hermitian or anti-Hermitian generally result in non-orthogonal eigenfunctions. Therefore, EOFs do not generally coincide with these mechanical modes, although there are notable situations when this does happen. North (1984) states that this occurs only if the two operators in question can commute, which in practice refers to the Hermitian and anti-Hermitian components of the operator.

In the case that they can commute, the eigenfunctions from the Hermitian and anti-Hermitian operator components are coincidental. As an example of a problem that does not meet these criteria, North (1984) cites the barotropic vorticity equation with latitude-dependent zonal winds. It is shown that the operator in this problem is not purely Hermitian or anti-Hermitian. The lack of an orthogonal basis and purely real eigenvalues suggests, according to North (1984), a possible mechanical instability. This appears to be the case, since EOFs of this system were previously noted to resemble instability modes more than standing waves. In more concrete terms, the symmetry between mechanical modes and EOFs only holds for systems with decay/growth modes when the decay/growth modes correspond to vibration or wave modes.

Results of frequency-dependent EOF analysis can be used to compare theoretical results to observational data to elucidate the sources of error, and as of 1993, frequency dependent EOFs were the preferred way to express natural variability (Kim and North, 1993). More specifically, Kim and North (1993) compare monthly surface temperature observations with a two-dimensional linear surface energy balance model. This model is coupled to a simple upwelling-diffusion ocean model. Notably, observational data shows a much lower variance in high-valued spatial modes than in that of the model data, and similarly, there is a discrepancy at low frequencies as well. Kim and North (1993) propose that these discrepancies can be explained by the relatively short period of available observational data. Kim and North (1993) implement a pattern-matching correlation coefficient and state that patterns that match the best between the EOFs and the theoretical orthogonal functions have some similarities with known oscillations. Using Monte Carlo simulations to artificially enhance the observational record length shows that sampling error does not sufficiently explain the relatively low amount of variance in higher

modes seen in the empirical analyses. Kim and North (1993) therefore suggest that a deficiency in the model physics is a more plausible explanation for the difference.

Cyclostationary EOFs are introduced in Kim et al. (1996). Until cyclostationary EOFs were defined, the time series analyzed by EOFs were required to be stationary to maintain orthogonality, which in the context of EOF analysis means that the statistical distribution parameters remain static for the entire time series. It is noted that over the annual cycle, variances and spatial autocorrelation distances do not, in fact, generally remain constant for the individual time series (Kim et al., 1996). Kim et al. (1996) use Bloch's theorem, a concept from solid-state physics, to inform an alternative formulation of the EOF kernel. This results in a nesting of modes: the cycles to be held as the periodic boundaries (e.g., the annual cycle) is held as the "outer" mode, while modes dependent on this outer mode are considered as "inner" modes. . In this form the kernel is a product of a Fourier function (which represent the outer modes) and a Bloch function (which represent the inner modes). These inner modes correspond to intra-annual modes and annual harmonics when the outer mode is the annual cycle. These inner modes are orthogonal to each other within each outer mode, but are not necessarily orthogonal to the inner modes of different outer modes. Imposing orthogonality is possible by simply defining the two bandwidths such that they do not overlap. Kim et al. (1996) use a theoretical model of an AR1 stochastic process with a periodic coefficient to show that oscillations included in such a process can be extracted with cyclostationary EOF analysis. This exact model is mirrored by the results of cyclostationary EOF analysis performed on a 100-year long set of surface temperature anomalies on a grid with $5^{\circ} \times 5^{\circ}$ spacing.

Since the previous research deals with a one-dimensional case and was already computationally expensive at the time; the two- and three-dimensional cases would, in

theory, be prohibitively expensive in terms of computing time (Kim and North, 1997). Therefore, a more efficient method is derived in Kim and North (1997). The first concept used to simplify calculation is to separate the nested sinusoidal signals from the outer modes, which can be accomplished when the nested modes are harmonizable. Each harmonizable nested mode contains a coefficient time series, representing the strength of the signal, but Kim and North (1997) note that this coefficient series is generally statistically stationary. Because of that, the covariance function of two coefficient series reduces to a single number that only depends on the lag. This results in the eigenvalue problem derived in Kim et al. (1996). Principal components are by definition uncorrelated at zero lag, but if it can be reasonably assumed that they are uncorrelated at all lags, as is usually the case, the eigenvalue problem's kernel can be simplified into a product of a Bloch function and a Fourier function, which are only functions of time, and not of lagged time as well. A space-time generalization is also offered for higher-dimensional applications of this method, wherein the only major difference is that the Bloch functions are now functions of space as well as time. A comparison of this new, computationally frugal method to the classic EOF method of Kim and North (1993) shows that the rate of eigenvalue decay is somewhat lower, due to the fact that intra-annual modes are differentiated by cyclostationary analysis.

A comparison study of eight EOF techniques is offered by Kim and Wu (1999). In this study, Kim and Wu (1999) examine the following methods: 1) classic EOF; 2) rotated EOF; 3) complex EOF; 4) extended EOF; 5) periodically extended EOF; 6) principal oscillation patterns; 7) cyclostationary principal oscillation patterns; and 8) cyclostationary EOF.

Rotated EOFs refer to an EOF solution that is adjusted in an attempt to minimize the number of major EOFs (Kim and Wu, 1999). This is typically done with VARIMAX rotation, but rotation is essentially arbitrary, particularly for interacting modes (North et al., 1982). Even so, VARIMAX and similar rotations typically result in more statistically stable spatial patterns (Kim and Wu, 1999), and therefore may yield results that more closely resemble physical modes. The VARIMAX and QUARTIMAX are useful when preservation of orthogonality is desired, although oblique rotations like QUARTIMIN have also been used (Hannachi et al., 2007). Rotation can also be weighted according to the eigenvalues, however, this can lead to convergence issues in oblique rotations (Hannachi et al., 2007). Hannachi et al. (2007) therefore use a superior definition of simplicity for EOFs from Jolliffe et al. (2003), called the “least absolute shrinkage and selection operator” (LASSO), which is referred to as simplified EOF (SEOF) analysis. SEOF analysis incorporates EOF rotation, but also attempts to minimize non-zero loadings in the principal components (Jolliffe et al., 2003). Though not always as robust as REOFs, the simplicity of SEOFs tends to result in more interpretable EOFs (Hannachi et al., 2007).

Complex EOFs refer to an analysis that initially transforms the raw data into a complex space using the Hilbert transform, and then solves the EOF problem as usual (Horel, 1984; Kim and Wu, 1999). This method is also referred to as complex Hilbert EOF analysis, or HEOF (Hannachi et al., 2007). This enables the derivation of phase data in addition to amplitudes, which allows for the identification of propagation. The results of complex EOFs are actually identical to the results of an LSEOF, as can be seen in the DISCUSSION section.

The extended EOF analysis is essentially a classic EOF analysis in which temporal lag is included as an additional dimension (Kim and Wu, 1999). For each lag of interest, a

covariance matrix is derived from the data. These covariance matrices can then be incorporated into a single block matrix, which is then the covariance matrix to be used in the eigenvalue problem. This can be expanded upon by treating each block as periodic in time, which is helpful in deriving spatial patterns from periodically varying quantities (Kim and Wu, 1999).

Principal oscillation pattern analysis (POPs) is a spatial pattern analysis based on an AR1 time series model (Hasselmann, 1988; Kim and Wu, 1999), and cyclostationary POP analysis (CSPOPs) is the extension for periodic variables. Principal oscillation patterns (POPs) are a specific, linear case of Principal Interaction Patterns (PIPs), which denotes a pattern analysis of nonlinear, autoregressive-moving average (ARMA) processes (Hasselmann, 1988). The results of POP analysis present as complex conjugate pairs for complex eigenvalues, which represent damped oscillations, and individual real-valued patterns for purely real eigenvalues, which represent damped exponential modes. POPs differ from the results of complex Hilbert EOFs in that POPs are more integrated into the spectral structure, and therefore yield useful information about the relative strengths of different oscillation patterns at each frequency. Hasselmann (1988) states that EOFs only fully explain time series oscillations when the complex cross-spectral covariance matrix is used, and is either evaluated at every frequency band or at each frequency in the spectral window. Another difference is that POPs are not generally orthogonal, although they form a basis set that is a linear transformation of an equivalent orthogonal basis (Hasselmann, 1988).

Methods such as the Data Interpolating Empirical Orthogonal Function (DINEOF) analysis were created in response to geophysical data with spatial and temporal gaps, when these gaps impede certain types of analysis (Alvera-Azcárate, 2011; hereafter AA11).

The core method of DINEOF analysis originates from Beckers and Rixen (2003; hereafter BR03). BR03 provides a technique to interpolate gappy data, and thereby derive EOFs from the data. This study was motivated by gappy oceanographic data in particular. BR03 notes that the SVD method assumes that the data matrix is fully defined, and that in cases where it is not, techniques such as objective analysis and optimal interpolation exist for filling these gaps. The problem with these methods is that this requires information about the correlation function and the signal-to-noise ratio that the EOF analysis is supposed to be trying to solve for in the first place. The contemporary solution was to use the filling method or one of its derivative methods. Alternatively, the covariance matrix method can be used with SVD to circumvent the problems posed by gappy data entirely. However, BR03 notes that this technique can result in negative eigenvalues, which would imply that positive-valued eigenvalues are inflated.

BR03 instead proposes to use placeholder data (in this case, zeroes) for the individual data points in question, and to use the EOFs to reevaluate what the placeholder data should be. This can be accomplished by deducing how many relevant EOFs exist, and then using those EOFs to interpolate the EOF value at the point in question for each EOF. This is accomplished by iteratively simulating data at one point based on the first EOF, and when it converges, adding the second EOF to the simulation process, and iterating likewise down to the last significant EOF.

AA11 notes that DINEOF was previously found to be up to thirty times faster than optimal interpolation, with similar accuracy, when an efficient EOF solver (from Toumazou and Creteaux, 2001) was included.

While this method may work for individual points, it assumes that the major EOFs' power is very large relative to the individual point's contribution to it. Therefore, if a data

set is highly gappy, the placeholder values may begin to influence the EOFs themselves significantly. One possible way to circumvent this might be to iterate through the points in question, rather than evaluating them at once, but this iteration in addition to the iterations already required for EOF convergence may render this method unfeasible for large data sets, like from long term satellite records. Regardless, DINEOF can be used to produce relatively accurate snapshots of interpolated data at an instant in time. It could also be used to downscale existing data, connecting in the process large-scale EOFs to scales where they would otherwise be considered insignificant, and thus discarded from many analyses.

Additionally, DINEOF must solve the EOFs and interpolated data points simultaneously because, as above, the correlation functions are unknown, but LSDFT EOF analysis can decouple the EOF problem from the interpolation problem. This implies that starting with a priori data and iterating to convergence, as in DINEOF analysis, would be unnecessary. Therefore, classic interpolation techniques can theoretically be used to reconstruct data from LSDFT EOFs about as accurately as DINEOF. However, in the case of downscaling, DINEOF analysis would still be useful because LSDFT EOF analysis cannot yield information about spatial points with no observations. Since the correlation function at such points is therefore unknown, the EOF problem and interpolation problem become coupled again in this situation.

Least-squares EOF (LSEOF) analysis is introduced in Boyd et al. (1994), and is a satisfactory technique for data sets with relatively sparse gaps. This method uses least-squares analysis to estimate a missing data point's value within a water column. Boyd et al. (1994) create synthetic data sets where an increasing number of points are deleted and estimated according to the model they propose. This synthetically completed data set is then processed as usual to derive the EOFs. It is noted that the deviations from the true

values increases as more values are deleted, and this is especially true above and below the thermocline. In particular, there are significant deviations of 1-2°C from the true values near the surface, where there is more temperature structure and variation, when the topmost values are missing. Removing bottommost values results in errors as well, but they are generally an order of magnitude smaller.

Another approach to the gappy EOF problem is the recursively subtracted EOF analysis, as proposed by Taylor et al. (2013). This method essentially borrows the LSEOF method for gap filling, but solves for only one EOF at a time. Each EOF is then used to estimate the respective data contribution to the EOF at each point. Iterating through the major EOFs should capture most of the variance, and the EOF coefficient matrix can be accurately calculated. The reconstructed data field is simply the product of the EOFs and their coefficient matrices. An advantage over DINEOF is that this method does not require iteration to convergence; the reconstructed data is defined explicitly. Though DINEOF yields reconstructions that are consistently slightly more accurate than RSEOF, Taylor et al. (2013) note that for their analysis, about 400 DINEOF iterations were needed for convergence for 70 EOF modes, while RSEOF analysis only ever requires a single iteration per EOF mode. However, DINEOF and RSEOF can also be combined by creating a better first guess from RSEOF, and then iterating through DINEOF as usual. Both of these methods yield superior results to LSEOF-based data reconstructions, although the LSEOFs themselves are generally acceptable.

Since EOF analysis does not require that the actual spatial points are arranged in any particular way, it may also be possible to interpolate a more classical EOF analysis by Kreiging or some other similar spatial interpolation.

Since the distinctions between EOFs can be somewhat arbitrary, Brunet (1994) proposed to reimagine EOFs in terms of empirical normal modes, or ENMs. In this study, Brunet (1994) examines empirical data for baroclinic wave activity over 24 winters. Brunet (1994) begins with a standard quasigeostrophic model, and introduces the idea of wave activity density, which is a result of a simple perturbation model. From this, there can be two kinds of wave activity, pseudomomentum and pseudoenergy. Applying the pseudomomentum and pseudoenergy definitions to the quasigeostrophic model results in a useful description of quasigeostrophic wave activity. For pseudomomentum, a monotonic potential vorticity guarantees that pseudomomentum is sign-definite, so it is a valid indicator of wave activity as long as it is globally conserved, and is still true when excluding local sources and sinks when it is not conserved (Brunet, 1994). The expectation value of pseudoenergy is the total energy difference between the unperturbed and perturbed model states. Brunet (1994) notes that pseudomomentum and pseudoenergy are the only two quadratic densities that are limits of nonlinear conservation laws, considering a small amplitude approximation, which means that taken together, they satisfy a uniqueness criterion. This holds so long as the basic quasigeostrophic potential vorticity equation leaves x and t explicitly independent. Brunet (1994) provides a detailed description of the application of normal mode theory to derive a simple relation between zonal wave speed and pseudoenergy, pseudomomentum, and normal mode wavenumber.

Brunet (1994) goes on to explain that in conservative barotropic and baroclinic models, the zonal phase speed spectrum can be divided into two branches: a continuous spectrum and a discrete spectrum. Notably, the continuous spectrum is associated with transient behavior, while the discrete spectrum is associated with longer-term oscillations. Brunet (1994) specifies that the timescale boundary between these two temporal regimes

is approximately two weeks. These phase speeds can be used to specify a wave function. The resulting wave function can in turn be expanded in terms of normal modes and wave activity. Brunet (1994) defines covariance operators for pseudomomentum and pseudoenergy based on this wave function, either of which can be used in an EOF problem. Solving the pseudomomentum EOF problem should result in spatial patterns associated with the particular normal modes of zonal phase speed. This enables the distinction of continuous phase speed modes from discrete phase speed modes. Solving the pseudoenergy EOF problem, though not mentioned, should result in spatial distributions of the relative strengths of wave activity.

Brunet (1994) extends all of this to the nonlinear case by defining a four-dimensional wave vector, which can be related to the nonlinear pseudomomentum covariance operator. Although the pseudomomentum operator has sign-indefinite eigenvalues, the eigenvectors are orthogonal. The nonlinear problem is therefore analogous to the linear case, but is more complex due to the higher dimensionality of the expression used to describe the waves themselves. It is shown that these techniques do actually result in more natural and readily interpretable spatial patterns, and Brunet (1994) also discusses methods to detect resonance and coupling between modes.

The ENM method described above has been compared to EOFs to determine the relative merits (Brunet and Vautard, 1996; hereafter BV96). For each of four experiments, BV96 derive linear statistical models from the quasi-geostrophic primitive equations, using the first 2000 model days to spin-up the model coefficients, and the second 2000 day period as the experimental data set. BV96 choose to examine the first few modes of zonal pseudomomentum (as a function of latitude) of the upper troposphere, which should be subject to excitation from bursts of wave activity in the lower atmosphere. For the first

three modes, including both eastward and westward components, the excitation of the first four principal components as a function of frequency is examined as the model is forced by Gaussian-distributed geopotential height perturbations. These impulses are found to occur on timescales of about 25 days on average, which suggests to BV96 a maximum relaxation time of four weeks. The time-averaged zonal flow is used as the basic state of the linearized model.

The four experiments in BV96 are differentiated by the amplitude of the height field disturbance, the larger of which is hypothesized to result in stronger nonlinear excitations. The physical agreement between the small amplitude and large amplitude experiments and empirical observations are generally good, with the exceptions that the nonlinearity in the large amplitude experiments leads to different latitudinal distributions of each ENM mode, and that the small amplitude experiment yields more monochromatic spectral peaks. Highlighting the importance of the choice of basic state, choosing a solid-body rotation as the model basic state results in more monochromatic behavior in the small amplitude case, but no significant difference is found in the large amplitude case.

BV96 also include an analysis of EOF and ENM skill by creating 20-day forecasts based on both methods every 25 model days, and comparing the model results to the forecasts. These forecasts show that ENMs are more skillful with high-wavenumber modes, while at modes one and two, ENM skill is statistically undifferentiable from EOF results. EOFs perform better than ENMs at low lead times (1-3 days) but BV96 state that this is likely an artifact of the skill scoring method they used, and ENMs perform better at longer lead times, particularly when time-averaged zonal flow is used as the basic state. This holds even in the more non-linear cases, though the lead times are lower across the board.

1.2 Exact Nonequispaced Time Series Methods

Lomb (1976) was the first to propose an exact solution to the problem of applying a Fourier Transform to unevenly spaced data. This developed in response to time series of variable star observations, but was also noted as relevant to ground-based astronomical data more generally. At the time, FFTs and the newly developed Maximum Entropy Method were the typical Fourier analyses available (Lomb, 1976). Classic periodograms were used for unevenly spaced data, but they do not account for the uneven spacing.

Aliasing is a major problem for all these techniques when using unevenly spaced data, and in practice, no more than one period could be deduced at a time (Lomb, 1976). Known periodicities could then be subtracted from the raw time series, in a process called “pre-whitening,” and subsequent periods could be found in Fourier analyses of the resulting pre-whitened data. For a classic periodogram, another shortcoming is that spectral peaks will not necessarily occur at their actual frequency. Other contemporary methods offered only marginal improvements over the periodogram. For these reasons, a least-squares method was extended to unevenly spaced data.

Scargle (1982) offers a similar analysis, citing the rise of automatic data generation as well as the planetary detection problem as reasons for developing a more robust time series analysis than classic periodograms. Highly regular signals do not raise the difficulty of periodogram statistics significantly, so given the choice between classic periodograms and Scargle’s analysis, the latter choice should offer a relatively inexpensive improvement. Additionally, smoothing schemes used for the classic periodogram are equally applicable to both.

Scargle (1982) also offers new insight into the statistical distribution of the periodogram, noting that when the general case of the classic periodogram is considered,

the result is exactly the same as a least-squares analysis of sinusoidal curves. Additionally, this analysis retains a time-translation invariance that is otherwise lost in the classic periodogram. This invariance has the effect of preventing spectral peaks from shifting from their true position.

The assumptions in Scargle (1982) include the idea that the time series is normally distributed, with zero mean and constant variance. Therefore, at this early stage of research into non-equispaced spectral analysis, cyclostationary and nonstationary time series could not be accounted for. However, simple detrending techniques are often sufficient to derive stationary time series from raw data. Indeed, Scargle (1982) demonstrates that least-squares sine curve fitting is equivalent to “folding” techniques frequently used in the astronomy community.

In retrospect, the techniques developed by Lomb (1976) and Scargle (1982) were recognized to be essentially identical, and were from then on known as the Lomb-Scargle periodogram (LSP).

At this stage however, the statistical confidence intervals were not immediately obvious. A “false alarm probability” had been prescribed (Scargle, 1982), but questions remained as to whether it utilized the correct normalization (Horne and Baliunas, 1986). Horne and Baliunas (1986) later established that, in fact, the variance of the LSP’s noise had an exponential distribution that was not captured in the false alarm probability normalization. Instead, they showed that the periodogram should be normalized by the total variance of the data to retrieve the correct false alarm probability (Horne and Baliunas, 1986).

Expanding the scope of non-equispaced time series analysis, Scargle introduces a complex formulation of the LSP. Since this complex formulation denotes a true discrete

Fourier transform, I choose to refer to it as the Lomb-Scargle Discrete Fourier Transform (LSDFT) to distinguish it from the LSP as well as other discrete Fourier transforms. Scargle (1989) notes that the inverse of the LSDFT need not take on a special formulation, so standard inverse Fourier transforms will suffice. Scargle (1989) demonstrates the effects of various maximum frequencies for an FFT version of the inverse LSDFT. Sample peaks are expressed as δ -like spikes for maximum frequencies higher than the standard choice for FFTs, while maximum frequencies lower than the standard choice result in a smoothing effect.

The complex formulation of the LSDFT creates a few numerical problems. First, the transform and the formula for the time-invariance constant introduced in Scargle (1982) are undefined when the frequency is zero. This is easily resolved by using a limit formulation. Second, the transform's complex term is undefined at the Nyquist frequency for equispaced data. This is resolved in Scargle (1989), but for non-equispaced data, this does not pose a problem. Third, the $\sqrt{\sin^2}$ and $\sqrt{\cos^2}$ terms in the transform and the *atan* function in the time-invariance constant each introduce a sign ambiguity. It is possible to resolve this by imposing continuity by reintroducing the appropriate signs. In practice, it is also possible to simply ignore it, since the results of imposing continuity are essentially the same. Although it is not clarified in Scargle (1989), the resolution of the sign ambiguity appears to come from the exponential term introduced therein, rather than just from a "cancellation" of the sign ambiguities of the trigonometric functions. This sign ambiguity is also noted later in Schulz (1997). Finally, large frequency values can create numbers too large for computers to handle during calculation, but this can be resolved by using integer multiples of 2π .

Lastly, Scargle (1989) adapts the autocorrelation and cross-correlation functions for non-equispaced data using the LSDFT, and demonstrates these using artificial and real astronomical data. Transforming two data sets (the same one for autocorrelation), multiplying one by the complex conjugate of the other, and taking the inverse transform yields the correlation function as a function of temporal lag. By using these transforms, the sampling times for the two series can be essentially arbitrary, as long as they overlap in phase-space.

To obtain more precise results, Schulz and Stattegger (1997; hereafter SS97) employed Welch's Overlapped Segment Averaging (WOSA) method in a time series analysis software package called "SPECTRUM." This research was motivated by unevenly sampled data in paleoclimatic time series. Schulz's approach does not address all the issues of periodogram biases, but it does offer a much-improved signal-to-noise ratio. SS97 created this package in response to the inability of the Blackman-Tukey method to handle non-equispaced data directly, which results in a redder spectrum than would otherwise be expected. Applying Welch's method to the Blackman-Tukey technique results in the additional problems that identical sampling times are required for both time series, relevant in the case of cross-spectra, and the fact that interpolated data are not statistically independent, which may result in significant biases.

SS97 also offer a relatively conservative scheme for adapting the WOSA method to bivariate time series. Though the basic formulae are straightforwardly derived from bivariate spectral analysis, calculating the average sampling interval and the fundamental frequency are both subject to some interpretation. Although SS97 choose the larger of the two available values, respectively, it may also be valid to choose the geometric means.

SS97 expand the analysis of bivariate time series by focusing on coherency, but although they dismiss the cross-spectrum itself as relatively unimportant, for the purpose of the research herein, the cross-spectrum is actually of greater importance than coherency. SS97 note that the complex-valued nature of the cross-spectrum entails a complicated statistical distribution—namely, the complex Wishart distribution—and describes the statistical significance of coherency. Since the partition of variance and its significance in EOF analysis can generally be described by North’s Rule of Thumb (North et al., 1982), it is not clear if it is necessary to propagate analysis of statistical significance, although doing so would most likely yield more accurate results.

Conveniently, Welch’s method deals with the individual transformed time series segments, rather than time series itself, and is therefore abstracted from the data enough so that it applies regardless of which technique is used to derive the transformed segments. This offers the possibility of using the LSP as the core Fourier analysis, as used in SPECTRUM, or the LSDFT for complex-valued results, or even some other form of spectral analysis.

To examine the effects that interpolation would otherwise have on the resulting Fourier analyses, SS97 compare the results of WOSA analysis with the LSDFT to WOSA analyses of data after using linear, cubic-spline, and Akima sub-spline interpolation for a given AR1 time series. Although all the peak amplitudes are collocated, interpolation results in dramatic variance losses: 54% for linear, 33% for cubic-spline, and 46% for Akima sub-spline.

Since paleoclimatic variables are typically autoregressive, it makes sense to examine the performance of statistical significance tests under different noise conditions (SS97). It appears that Siegel’s test, which can detect up to three periodicities, performs

normally for unevenly sampled data when white noise is present. However, a red noise situation can result in spurious peaks for which the null hypothesis is rejected, which implies that harmonic analyses such as Siegel's test should not be taken at face value for autoregressive time series.

To address the biases usually associated with meteorological and climatological time series, Mudelsee (2002) and Schulz and Mudelsee (2002; hereafter SM02) together offer a combination of techniques to evaluate the AR1 spectra associated with such time series.

An accurate estimate of the persistence associated with a given time series needs to be calculated first (Mudelsee, 2002). Previously, time series had been parameterized in terms of recursion plus errors, wherein a single constant represents the persistence time (Robinson, 1977). A least-squares method can be used to estimate this persistence time. However, there is a known bias when using equally sampled data, and a similar bias for unevenly sampled data, so Mudelsee (2002) provides a prescription for correcting these biases as well.

With the persistence time known, it is now theoretically possible to perform Monte Carlo simulations, as is needed to find the AR1 spectra (SM02). The theoretical AR1 noise spectrum can be calculated relatively easily, but the spectrum still remains biased. Assuming that the time series has a known distribution, the statistical moments can be used to generate synthetic time series. SM02 uses these synthetic time series to correct a bias caused by the fact that LSDFT components may be correlated, which depends on the variations in sampling unevenness, and is therefore inherent in LSDFT spectra.

Bretthorst and Feigelson (2003; hereafter BF03) generalize the periodogram using Bayesian probability theory, noting that the LSP, among a few others, is a particular case of

this generalization. This analysis was developed to accommodate quadrature data, in which both the real and imaginary components of the data set are sampled, and the imaginary part is sampled at 90 degrees out of phase from the real part. The spectral analysis formulae resulting from Bayesian analysis allow the imaginary and real components to be sampled independently. However, the orthogonality requirement can be satisfied by using identical sampling times for both components; otherwise, it will have to be satisfied by generalizing the LSP, which in practice means applying a formula (provided in BF03) to solve for the phase. This generalized periodogram can also account for secular, time-dependent trends, which broadens the possible uses of the LSP to time series in which the secular process is known. BF03 also gives examples of other particular cases that can easily be derived from the generalized periodogram. Though BF03 assumes a Gaussian noise profile, alternate noise models can be used by rederiving part of the analysis, but the fundamental scheme is unaffected.

It should be noted that, as per BF03, probability theory cannot literally generalize a Fourier analysis. However, Bretthorst and Feigelson (2003) found that, while examining the optimal technique according to Bayesian probability theory, the sufficient statistic turns out to be a generalization of several Fourier analysis techniques, including the LSP. However, they are only sufficient statistics in the case of single-sinusoid analyses, and are never sufficient statistics in the case of multiple-sinusoid analyses, although there are options for improving upon using LSPs for multiple-sinusoid analyses.

Bayesian analysis (BF03) makes one interpretation of the final probability distribution viable: the width of a frequency peak can be summarized by a mean and a standard deviation, denoting the probabilities that a frequency peak lies within a particular range. This would contrast with the possibility that broadened frequency peaks may imply

stochastic variation of the particular frequency within a certain range, or simply be a result of spectral leakage. From a purely mathematical point of view, it may be impossible to tell the difference between the three.

BF03 continues by discussing aliasing within the context of LSP analysis. Though the sampling time intervals are irregular, a time interval satisfying a prescribed condition can be determined, and by basic aliasing theory, spectral data falling outside the resulting bandwidth is by definition aliased. Additionally, it is noted that regularly sampled data actually have the smallest possible bandwidth; according to the scheme used to derive bandwidth in the general case, uneven sampling will always result in a larger range of unaliased Fourier analysis.

Mathias et al. (2004) provide an analysis that, in addition to providing a least-squares analysis of unevenly sampled data, includes a scheme for meaningful phase data. This particular study includes example analyses of paleoclimatic data from an Antarctic ice core, as well as tick data from the London stock exchange. By formulating the Fourier analysis in matrix form, it is demonstrated that such a Fourier analysis yields essentially the same results even when the matrix is rotated in phase-space. In the context of time series, this matrix rotation is equivalent to a shift in sampling times, which explains why this should be the case. Notably, this formulation can accommodate complex data, although it is not as robust as Bretthorst's (2003) analysis. This formulation of LSP analysis is ideal for the present study, because while the data used is purely real, the imaginary component of the Fourier analysis will be useful. Additionally, the ability to introduce phase shifts allows for the addition of a reference point to interpret the otherwise-arbitrary phases: in the diurnal case, it allows us to interpret these phases as times of day.

With the advent of powerful graphical processing units (or GPUs for short) in the 2000s, the calculation of massively parallelizable problems suddenly became much more feasible. This is especially true since the development of CUDA (Nickolls et al., 2008) and OpenGL, which allow for easy interfacing between common programming languages like C or Fortran and the GPU itself, reducing the specialized knowledge requirements for GPU programming. GPU-based algorithms have gradually been entering many scientific realms since then. Motivated by astronomical time series, Townsend (2010) opts to develop the classic LSP algorithm into a GPU script, rather than use FFT-based approximations. This algorithm is designated CULSP. Townsend (2010) finds that on a per-core basis, CUDA performs approximately as well as CPU calculations, indicating that there is negligible linguistic overhead with CUDA programs.

1.3 Approximate Nonequispaced Time Series Methods

The LSP was (and is) still remarkably slow compared to FFT algorithms. Press and Rybicki (1989) presented a new approach using FFTs that reduces the number of calculations from $O(N^2)$ to $O(N \ln(N))$, the theoretical scaling of the FFT's speed, modulo some factor. This algorithm “extirpolates,” that is to say, reverse interpolates, unevenly spaced data. Whereas in interpolation, the value at an arbitrary point is defined in terms of the values of several gridded points, in extirpolation, the values at those gridded points are defined in terms of the arbitrary points' values. For a grid large enough to oversample a particular minimum wavelength a specified number of times, the extirpolated data and its point weights can be processed by an FFT algorithm. From there, it is a matter of arithmetic to derive a solution to the LSP.

This technique was quickly adopted by others in the astronomy community (e.g., Scargle 1989), and has been extended to many applications, though the basic algorithm remained largely unaltered as research progressed.

The FFT adaptation of the LSP pioneered by Press and Rybicki (1989) is expanded upon by LeRoy (2012). This improvement is motivated by astronomical data from the CoRoT and Kepler satellites. Exact nonequispaced FFT (or NFFT) methods (e.g., Dutt and Rokhlin 1993; Steidl 1998) and GPU-based classic LSDFT algorithms are noted as viable alternatives. LeRoy (2012) is able to bypass the extrapolation algorithm necessary in Press and Rybicki (1989) by making use of open source NFFT software (Keiner et al., 2009). Reducing the complexity of the LSP numerator terms and weeding out the duplicate Fourier components enables LeRoy's (2012) algorithm to achieve speeds that are about an order of magnitude faster than Press and Rybicki's (1989), and about five times as fast as GPU-based classic algorithms. Though the solution here is inexact by definition, it is apparently able to manage these speed increases without losing a significant amount of accuracy.

1.4 The Statistics of Rainfall

The statistical distribution of rain rate is an active area of research. Since the AR1 simulation (Schulz and Mudelsee 2002) depends heavily on realistic simulated data, it becomes important to choose a reasonable statistical distribution. The distribution in turn is usually considered to be a mixed distribution, composed of an atom at zero rain rate (denoting dry observations) and a more typical statistical distribution at non-zero values (denoting the rain rates on wet observations). The existence of an atom at zero appears to be a consensus opinion, but the wet observation distribution appears to be an unsettled

question. The lognormal (e.g., Kedem et al., 1990) and 3-parameter gamma (also known as Pearson Type III) distributions (e.g., Husak et al., 2007; Amirataee and Montaseri 2013) appear to be the most successful models, and are the most frequently used. The success of these models may depend on time of year (Amirataee and Montaseri 2013) and sampling frequency (Sharma and Singh 2010). They may also perform better when incorporated into an autoregressive model (Şarlak and Şorman, 2007). Other models are also used occasionally, and can outperform those two in certain contexts (e.g., Aksoy, 2000; Hanson and Vogel, 2008; Amirataee and Montaseri, 2013). Alternatively, random cascades can also be used to create realistic rainfall distributions (Menabde and Sivapalan, 2000).

For the purposes of scaling power spectra as per SM02, it may not be necessary to use non-Gaussian statistics for the wet observation distribution. The Central Limit Theorem allows for the possibility that if a small number of non-Gaussian processes can be isolated, subtracting these from the raw data will yield data that more closely fits a normal distribution.

1.5 Computational Aspects

The fact that the fundamental problem presented here is potentially massively parallelizable presents an ideal test case for comparing performance of vector-oriented and GPU-based programming paradigms with respect to geoscience problems. The R programming language (R Core Team, 2015) is typically thought to be extremely slow compared to languages such as C and Fortran (e.g., Morandat et al., 2012; Aruoba and Fernández-Villaverde, 2014). Aruoba and Fernández-Villaverde (2014) find that R is 475 to 491 times slower than C++ when uncompiled, and 243 to 282 times slower when compiled, although it is noted that using Rcpp for C++-style programming results in speeds only

about 3.66 and 5.41 times slower. However, previous speed tests appear to have been performed with problems ill-suited to R, and by those who are not particularly fluent in R. Additionally, R now is packaged with a byte compiler, which can accelerate R scripts dramatically. In sum, there are too many variables that have not been accounted for in previous studies of R's speed to arrive at any particular conclusion.

2. METHODS

2.1 Empirical Orthogonal Functions

For a vector of time series $\mathbf{S}_{s \times t}$, where s is the number of spatial points and t is the number of temporal points arranged as

$$\mathbf{S}_{s \times t} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1t} \\ s_{21} & s_{22} & \cdots & s_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ s_{s1} & s_{s2} & \cdots & s_{st} \end{pmatrix}, \quad (2.1)$$

wherein the rows can be treated as individual time series, and the covariance matrix \mathbf{C} is defined as

$$\mathbf{C} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1j} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{i1} & \sigma_{i2} & \cdots & \sigma_{ij} \end{pmatrix} = \mathbf{S}\mathbf{S}^T. \quad (2.2)$$

The EOF eigenvalue problem is thus given as $\mathbf{A}\mathbf{C} = \lambda\mathbf{C}$ where \mathbf{A}_{nj} denotes the n^{th} EOF's j^{th} spatial element, and λ_n^2 denotes the n^{th} EOF's explained variance. Solving an evenly sampled EOF problem is fairly straightforward, but to solve the EOF problem directly with unevenly sampled data is quite difficult, if not impossible. The LSDFT offers some respite, as it is possible to relate covariance of each element to its respective transform through the cross-correlation theorem:

$$\sigma_{xy}(t_{lag}) = \int_0^\infty X(t)Y^*(t - t_{lag})dt = \int_0^\infty \tilde{X}(\omega)\tilde{Y}^*(\omega)e^{i\omega t_{lag}}d\omega \quad (2.3)$$

It should be noted that by using this relation, it becomes unnecessary to calculate the cross-spectral Lomb-Scargle transforms; only univariate spectra are needed. Additionally, while EOFs are typically calculated with $t_{lag} = 0$, it is possible to calculate EOFs for nonzero lags.

When a lag is applied, the result is called an “extended” LSDFT (Monahan et al., 1999). For a transformed cross spectrum of time series X and Y , denoted as $\widetilde{XY}(\omega)$, the relation is

$$\sigma_{xy}(t_{lag}) = \int_0^\infty \tilde{X}(\omega)\tilde{Y}^*(\omega)e^{i\omega t_{lag}}d\omega = \int_0^\infty |\widetilde{XY}(\omega)|^2 e^{i\omega t_{lag}}d\omega \quad (2.4)$$

Using the first integral (henceforth the “univariate LSEOF integral”) rather than the second (henceforth the “bivariate LSEOF integral”) will reduce the number of LSDFT calculations from order $O(N^2)$ to $O(N)$ plus a relatively small overhead for multiplication of the transformed time series. The workflow for each is outlined in Figure 2.

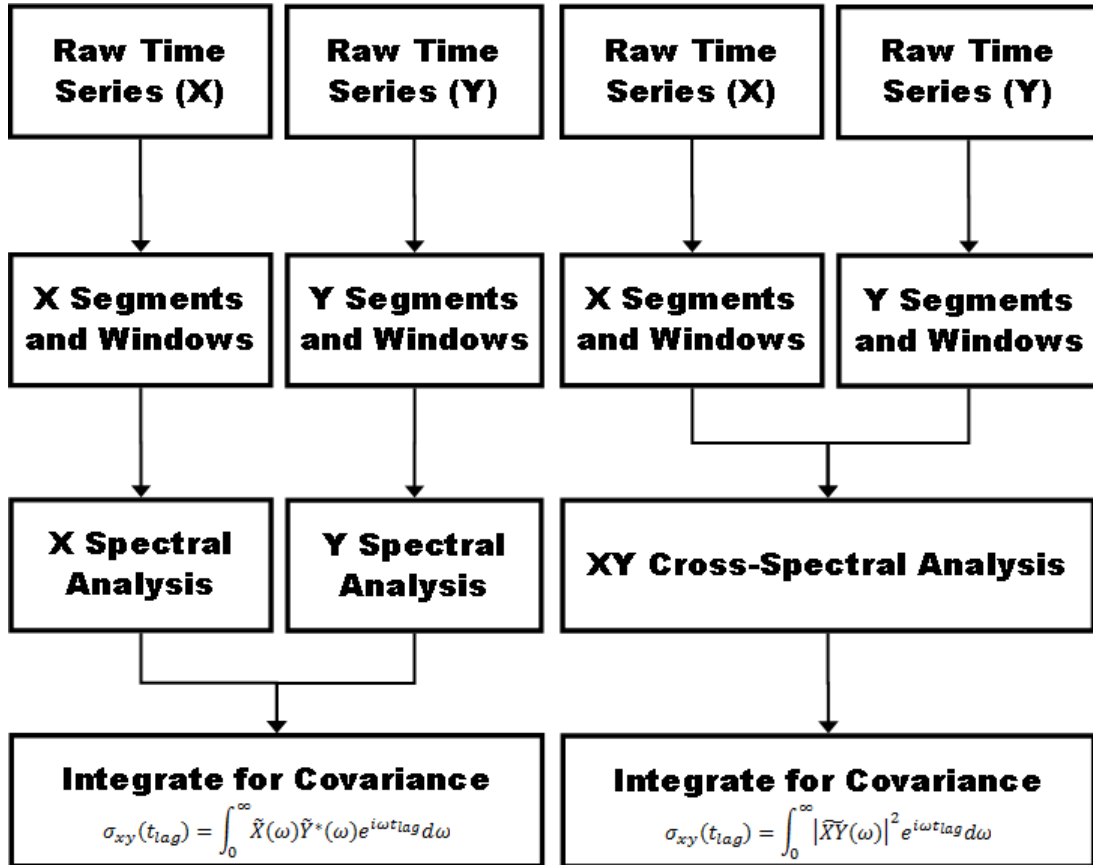


Figure 2: Workflows for calculating covariance using (left) the univariate integral and (right) the bivariate integral.

Examining the statistical significance of each EOF is made possible by North's Rule of Thumb (North et al., 1982). The variance confidence interval $\Delta\lambda^2$ is defined as

$$\Delta\lambda^2 = \lambda^2 \sqrt{2/N}, \quad (2.5)$$

where N is the number of independent samples. Although North's Rule of Thumb implicitly assumes a constant number of samples across the data field, a mean sample number can offer an estimate.

2.2 The Lomb-Scargle Discrete Fourier Transform

The periodogram of a time series is classically defined as $P(\omega) = \frac{1}{N} |\tilde{S}(\omega)|^2$, where $P(\omega)$ denotes the spectral power density and N_0 is the number of samples in the original time series. While useful for its simplicity, the modulus function in the periodogram ensures that half the spectral information from a complex-valued DFT will be lost. While the imaginary part is never strictly necessary, it will be useful for reducing the number of necessary computations further along in the process of creating a covariance matrix. For cross spectra, the periodogram can be generalized to $P_{1,2}(\omega) = \frac{1}{N} |\tilde{S}_1(\omega)| |\tilde{S}_2^*(\omega)|$, where N can be defined as $\sqrt{N_1 N_2}$, the geometric mean of the time series lengths. However, the normalized periodogram is defined with a constant inversely proportional to the total variance, as opposed to dividing by the number of observations (Horne and Baliunas, 1986).

The Lomb-Scargle periodogram as classically derived can be found in Scargle (1982), as well as numerous other papers on the topic. For evenly spaced data, t is simply the time between samples, but to preserve the "invariance of time translation," as Scargle (1982) phrases it, a constant τ must be defined as

$$\tau = \frac{1}{2\omega} \tan^{-1} \left(\frac{\sum_j \sin(2\omega t_j)}{\sum_j \cos(2\omega t_j)} \right), \quad (2.6)$$

which can then be used to augment t where it occurs in the basic periodogram equation.

The discretized form of the periodogram is therefore defined for spectra as

$$\begin{aligned} P(\omega) &= \frac{1}{2\sigma^2} |\tilde{S}(\omega)|^2 = \frac{1}{2\sigma^2} \left| \sum_{j=1}^{N_0} (S(t_j) - \bar{S}) e^{-i\omega(t_j - \tau)} \right|^2 \\ &= \frac{1}{2\sigma^2} \left(\frac{\left(\sum_j (S_j - \bar{S}) \cos(\omega(t_j - \tau)) \right)^2}{\sum_j \cos^2(\omega(t_j - \tau))} + \frac{\left(\sum_j (S_j - \bar{S}) \sin(\omega(t_j - \tau)) \right)^2}{\sum_j \sin^2(\omega(t_j - \tau))} \right), \end{aligned} \quad (2.7)$$

and more generally for cross-spectra as

$$\begin{aligned} P_{1,2}(\omega) &= \frac{1}{2\sigma_1\sigma_2} |\tilde{S}_1(\omega)| |\tilde{S}_2^*(\omega)| \\ &= \frac{1}{2\sigma_1\sigma_2} \left| \sum_{j=1}^{N_0} (S_1(t_{1,j}) - \bar{S}_1) e^{-i\omega(t_{1,j} - \tau_1)} \right| * \left| \sum_{j=1}^{N_0} (S_2^*(t_{2,j}) - \bar{S}_2^*) e^{i\omega(t_{2,j} - \tau_2)} \right| \\ &= \frac{1}{2\sigma_1\sigma_2} \left(\frac{\sum_j (S_{1,j} - \bar{S}_1) \cos(\omega(t_{1,j} - \tau_1))}{\sqrt{\sum_j \cos^2(\omega(t_{1,j} - \tau_1))}} + \frac{\sum_j (S_{1,j} - \bar{S}_1) \sin(\omega(t_{1,j} - \tau_1))}{\sqrt{\sum_j \sin^2(\omega(t_{1,j} - \tau_1))}} \right) \\ &\quad + \frac{1}{2\sigma_1\sigma_2} \left(\frac{\sum_j (S_{2,j}^* - \bar{S}_2^*) \cos(\omega(t_{2,j} - \tau_2))}{\sqrt{\sum_j \cos^2(\omega(t_{2,j} - \tau_2))}} - \frac{\sum_j (S_{2,j}^* - \bar{S}_2^*) \sin(\omega(t_{2,j} - \tau_2))}{\sqrt{\sum_j \sin^2(\omega(t_{2,j} - \tau_2))}} \right). \end{aligned} \quad (2.8)$$

However, the nature of the technique used herein for solving the EOF problem suggests the use of the full, complex-valued solution. Considering the classical discrete Fourier transform

$$\tilde{S}(\omega) = \frac{1}{\sqrt{2\sigma^2}} \sum_j (S_j - \bar{S}) \cos(\omega t_j) + i(S_j - \bar{S}) \sin(\omega t_j), \quad (2.9)$$

and its associated periodogram,

$$P(\omega) = \frac{1}{2\sigma^2} \left(\left(\sum_j (S_j - \bar{S}) \cos(\omega t_j) \right)^2 + \left(\sum_j (S_j - \bar{S}) \sin(\omega t_j) \right)^2 \right), \quad (2.10)$$

respectively, and the formula for the Lomb-Scargle periodogram above, I argue by comparison that the logical formulation for the complex-valued Lomb-Scargle discrete Fourier transform (LSDFT) is the following:

$$\tilde{S}(\omega) = \frac{1}{\sqrt{2}\sigma^2} \sum_j \left(\frac{(S_j - \bar{S}) \cos(\omega(t_j - \tau))}{\sqrt{\sum_j \cos^2(\omega(t_j - \tau))}} + \frac{i(S_j - \bar{S}) \sin(\omega(t_j - \tau))}{\sqrt{\sum_j \sin^2(\omega(t_j - \tau))}} \right). \quad (2.11)$$

This formulation is essentially the same as that of Scargle (1989) and Mathias et al. (2004).

Superior implementations of the LSDFT are more complicated. Using Welch's Overlapped Segment Averaging (WOSA) technique, it is possible to dramatically improve the signal-to-noise ratio (Schulz, 1997). Note that for this technique, σ^2 is omitted from the periodogram definition itself. The cross-spectrum from WOSA (\hat{G}_{xy}), is calculated by

$$W_{xy}(\omega_j) = \frac{2}{n_{50} \sqrt{N_{seg}^{(x)} N_{seg}^{(y)} \Delta f_{xy}}} \sum_{n=1}^{n_{50}} \left(H(\tilde{X}_n(\omega_j) \{I_n\}) H(\tilde{Y}_n^*(\omega_j) \{I_n\}) \right)^2 \quad (2.12)$$

$$\hat{G}_{xy}(\omega_j) = \frac{\sigma_x \sigma_y}{\Delta f \sum_j W_{xy}(\omega_j)} W_{xy}(\omega_j), \quad (2.13)$$

where I denotes the time interval

$$I = \left\{ \frac{t_{max}(n-1)}{n_{50}+1} \leq t \leq \frac{t_{max}(n+1)}{n_{50}+1} \right\}, \quad (2.14)$$

and where H is a windowing function. In this study, the Hamming window is used, which is defined as

$$H(S_j) = \frac{(S_j - \bar{S}) \left(\alpha - \beta \cos \left(2\pi \left(\frac{t_j - t_{min}}{t_{max} - t_{min}} \right) \right) \right)}{\sqrt{\frac{1}{N_{seg}} \sum_j \left(\alpha - \beta \cos \left(2\pi \left(\frac{t_j - t_{min}}{t_{max} - t_{min}} \right) \right) \right)^2}}, \quad (2.15)$$

where $\alpha = 0.53836$ and $\beta = 1 - \alpha$. It is important to note that this windowing function's bounds are defined by minimum and maximum time values. While in the evenly-spaced case it is inconsequential, conflation of windowing the increment difference (as is common) and windowing the time difference will result in an incorrect window function, skewed in the time domain depending on the temporal differences between individual points. Therefore, while it is somewhat non-standard, I choose the above definition. The same observation should be considered for other windowing functions as well.

A caveat of using a complex form of the LSDFT is that Schulz (1997) offers no method to deal with complex data. Since phase data is defined on a circular domain, segment averaging will require the computation of circular means for phase. Phase is typically defined as

$$\varphi = \text{atan2}(\text{Im}(X_n), \text{Re}(X_n)). \quad (2.16)$$

Therefore, the phase of the LSDFT is simply

$$\varphi(\tilde{S}(\omega)) = \text{atan2} \left(\sum_j \left(\frac{(S_j - \bar{S}) \sin(\omega(t_j - \tau))}{\sqrt{\sum_j \sin^2(\omega(t_j - \tau))}} \right), \sum_j \left(\frac{(S_j - \bar{S}) \cos(\omega(t_j - \tau))}{\sqrt{\sum_j \cos^2(\omega(t_j - \tau))}} \right) \right), \quad (2.17)$$

and in the bivariate case,

$$\varphi(\tilde{S}(\omega)) = \text{atan2}(a, b) \quad (2.18)$$

$$a = \sum_j \left(\frac{(S_{1,j} - \bar{S}_1) \sin(\omega(t_{1,j} - \tau_1))}{\sqrt{\sum_j \sin^2(\omega(t_{1,j} - \tau_1))}} - \frac{(S_{2,j}^* - \bar{S}_2^*) \sin(\omega(t_{2,j} - \tau_2))}{\sqrt{\sum_j \sin^2(\omega(t_{2,j} - \tau_2))}} \right) \quad (2.19)$$

$$b = \sum_j \left(\frac{(S_{1,j} - \bar{S}_1) \cos(\omega(t_{1,j} - \tau_1))}{\sqrt{\sum_j \cos^2(\omega(t_{1,j} - \tau_1))}} + \frac{(S_{2,j}^* - \bar{S}_2^*) \cos(\omega(t_{2,j} - \tau_2))}{\sqrt{\sum_j \cos^2(\omega(t_{2,j} - \tau_2))}} \right). \quad (2.20)$$

For the WOSA algorithm, Schulz (1997) offers the formula

$$\varphi = \text{atan2}\left(\text{Im}\left(\hat{G}_{xy}(\omega_j)\right), \text{Re}\left(\hat{G}_{xy}(\omega_j)\right)\right) \quad (2.21)$$

for calculating the phase, but this cannot be the full answer, because in the univariate case, the time series complex conjugate pair multiplication in Equation (2.12) implies that for every ω_j , $\text{Im}\left(\hat{G}_{xy}(\omega_j)\right) = 0$. This further implies that there can be no phase differences across a spatial field, which is obviously not the case. The “phase” as described in Schulz (1997) therefore only describes the phase difference between the two time series in a bivariate analysis, not the actual phase per se.

For reasons that have been explained in Section 2.1, the EOF problem is computed much more quickly when it is expressed in terms of univariate time series as opposed to bivariate time series. Because of that, we must seek an alternative way to calculate the phase such that it gives meaningful results in the univariate case as well. The mean phase is required in the WOSA algorithm, but it is not so straightforward to calculate. Because the phase is a periodic quantity, the circular mean must be used instead of the arithmetic mean. The circular mean is defined as

$$\bar{\varphi} = \text{atan2}(\overline{\sin(\varphi_n)}, \overline{\cos(\varphi_n)}), \quad (2.22)$$

which results in the following formula for WOSA phase:

$$\bar{\varphi}(\hat{G}_{xy}(\omega_j)) = \text{atan2}\left(\overline{\sin\left(\varphi\left(C_n(\omega_j)\right)\right)}, \overline{\cos\left(\varphi\left(C_n(\omega_j)\right)\right)}\right), \quad (2.23)$$

where

$$C_n(\omega_j) = H(\tilde{X}_n(\omega_j)\{I_n\})H(\tilde{Y}_n^*(\omega_j)\{I_n\}). \quad (2.24)$$

At this point, the spectrum remains red, and for most atmospheric purposes, the baseline can be successfully modeled as an autoregressive process of order one (AR1). This can be mitigated through the use of a Monte-Carlo simulation based on statistics from WOSA results (SM02). The general formula for an AR1 process is

$$X_n(t) = \rho X_{n-1}(t) + \varepsilon_n. \quad (2.25)$$

This generalization assumes that X is equispaced though, which yields a constant value of ρ . A more robust prescription is necessary, and Robinson (1977) offers the necessary robustness, formulating the AR1 process as

$$X_n(t) = X_{n-1}(t)e^{-(t_n - t_{n-1})/\tau} + \varepsilon_n. \quad (2.26)$$

In this context, τ refers to a persistence factor, and can be found through the TAUEST algorithm outlined in Mudelsee (2002). When using multiple segments in a WOSA algorithm, SM02 uses an average of the segment persistence factors. The error term ε is assumed to be some random process, and for the purposes of SM02 was assumed to be normally distributed with a mean of zero and a variance of $1 - e^{-2(t_n - t_{n-1})/\tau}$. This is derived by taking the moments of Equation (2.26) and solving:

$$\overline{X_n} = \rho \overline{X_{n-1}} + \overline{\varepsilon_n} \quad (2.27)$$

$$\mu = \rho \mu + \mu_\varepsilon \quad (2.28)$$

$$\mu_\varepsilon = \mu(1 - \rho) \quad (2.29)$$

$$\text{Var}(X_n) = \text{Var}(\rho X_{n-1}) + \text{Var}(\varepsilon_n) \quad (2.30)$$

$$\sigma^2 = \rho^2 \sigma^2 + \sigma_\varepsilon^2 \quad (2.31)$$

$$\sigma_\varepsilon^2 = \sigma^2(1 - \rho^2) \quad (2.32)$$

The above set of equations applies for a normal distribution. However, so long as the moments can be calculated and an appropriate random number generator can be coded, there is no hard limit to the types of random processes that could be used instead. The theoretical value of a purely random AR1 time series' Fourier transform (G_{rr}) is

$$G_{rr}(\omega_i) = G_0 * \frac{1 - \rho^2}{1 - 2\rho \cos(\pi\omega_i/\omega_{Nyq}) + \rho^2} \quad (2.33)$$

$$\rho = e^{-\frac{t_N - t_1}{N-1} * \frac{1}{\tau}} \quad (2.34)$$

$$G_0 = \frac{\int_{\omega_{min}}^{\omega_{max}} G_{rr} d\omega}{\int_{\omega_{min}}^{\omega_{max}} G_{xy} d\omega} \quad (2.35)$$

However, this must be corrected for bias. This correction factor c can be calculated by performing a spectral analysis of N_{sim} synthetic time series. The resulting spectra $\hat{G}_{rr}(\omega_i)$ are then averaged for each frequency.

$$c = \frac{\frac{1}{N_{sim}} \sum_{m=1}^{N_{sim}} \hat{G}_{rr}^{(m)}(\omega_i)}{G_{rr}(\omega_i)} \quad (2.36)$$

The final, corrected spectrum G'_{xy} can then be calculated from the uncorrected spectrum:

$$G'_{xy} = \frac{G_{xy}}{c} \quad (2.37)$$

This method for AR1 bias spectrum calculation, like the others, does not include a prescription for complex spectral results. Since SM02 applies primarily to periodograms, the method here could be used to enhance the amplitude results as long as the complex components of the spectra are formulated in modulus-argument form. This does not, however, provide a solution to the argument component, and in fact, simply applying the same technique to the argument (with circular averaging, as with WOSA) will introduce a

spurious randomization of phases. Therefore, an alternative statistical model that can incorporate appropriate phase information must be developed in order for the SM02 method to yield complex results with similar precision improvements both in amplitude and phase. Since that is beyond the scope of this research, I choose instead to simply propagate the phase data from the original WOSA algorithm, and combine it with the new amplitude spectrum from the AR1 Monte Carlo simulation.

For the purposes of this study, the false alarm probability will not be particularly helpful, since it deals primarily with relatively small time series (Horne and Baliunas, 1986). The exponential dependence on the number of independent observations ensures that false alarm probabilities rapidly collapse toward zero as the size of the time series increases.

To demonstrate the robustness of these algorithms in dealing with large gaps, 15000 contiguous samples from TRMM are omitted from the data set, and the resulting data is processed with WOSA LSDFT analysis.

2.3 Statistical Distribution of Rainfall

The present techniques are valid for stationary, normally-distributed processes, and are capable of handling AR1 processes using Monte Carlo simulations (SM02). When alternative distributions are necessary, the algorithm thus far must be extended to these non-Gaussian distributions. In particular, reasonable Monte Carlo simulations will require a robust statistical model.

In particular, I consider the case of 3-hourly rainfall data collected (3B42) from the Tropical Rainfall Measurement Mission (TRMM), which is freely available from the National Air and Space Administration (NASA). The statistical distribution of precipitation

is an active area of research, and recent findings indicate that there may not be a single, unified answer. The best model for a given location may change based on sampling rate, and varying meteorological predispositions may impact different locations. For a probability of rainfall r and based on prior knowledge, the conditional probability distribution function $P(r, x)$ can be described as

$$P(r, x) = r\delta_{[q-r]}g(x), \quad (2.38)$$

where $g(x)$ represents a mathematically arbitrary choice of statistical distributions, δ_x is the Kronecker delta function, and values of q are randomly generated realizations of the empirical quantile function of x . Values of q range from zero to one, and since it is already distributed with respect to x , q can be modeled by a uniformly distributed random number generator.

The conditional cumulative distribution $C(r, x)$ is then given as

$$C(r, x) = (1 - r)\delta_{[q-r]} + r\Theta([q - r])\dot{\xi}(x_n), \quad (2.39)$$

where $\dot{\xi}(x_n)$ is the cumulative distribution function corresponding to $g(x)$, and the Heaviside step function $\Theta(x)$ represents the relatively less common definition of

$$\Theta(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases}. \quad (2.40)$$

Since rainfall is an autoregressive process, it will be include a term representing autoregression. As before,

$$X_n(t) = X_{n-1}(t)e^{-(t_n - t_{n-1})/\tau} + \varepsilon_n. \quad (2.41)$$

The error term ε_n in the case of rainfall is a $P(r, x)$ -distributed random variable. Combining these two distributions requires the prioritization of the two conditions introduced previously. I choose to give priority to the autoregressional portion, resulting in a model for autoregressive, g -distributed rainfall

$$X_n = X_{n-1}e^{-(t_n-t_{n-1})/\tau} + r_n\delta_{[q_n-r_n]}g(x_n). \quad (2.42)$$

This model allows for variable probabilities of rainfall, but for stationary time series, it should remain constant. Alternatively, if priority is given to the probability of rainfall,

$$X_n = r_{n-1}\delta_{[q_{n-1}-r_{n-1}]}g(x_{n-1})e^{-(t_n-t_{n-1})/\tau} + r_n\delta_{[q_n-r_n]}g(x_n), \quad (2.43)$$

which is equivalent to Equation (2.42).

The parameters of the chosen distribution for rainfall need to be optimized for each time series after subtracting the autoregressive term from $X_n(t)$. With all the model parameters known, artificial data can be generated. To do so, it is necessary to create a cumulative form of the rainfall distribution function. Since the probability distribution is already known, the relation

$$F_n(t) = \int_{-\infty}^{x_n} X_n(t) dx_n \quad (2.44)$$

can be used to derive the cumulative distribution function:

$$F_n(t) = \int_{-\infty}^{x_n} X_{n-1}e^{-(t_n-t_{n-1})/\tau} dx_n + \int_{-\infty}^{x_n} \delta_{[q_n-r_n]}g(x_n) dx_n \quad (2.45)$$

$$F_n(t) = X_{n-1}x_ne^{-(t_n-t_{n-1})/\tau} + (1-r_n)\delta_{[q_n-r_n]} + r_n\theta([q_n-r_n])\dot{\xi}(x_n) \quad (2.46)$$

For a uniformly distributed quantile function, an appropriate randomization scheme can begin by assigning a uniformly distributed random number to q_n . From there, the value of x_n can be estimated.

Finally, with the distribution parameters known, the correct LSDFT variance (σ^2) is usually easy to calculate. The final normalization will depend on where this variance definition is applied, however. By the normalization provided by Horne and Baliunas (1986), the correct place to apply this relation is at the periodogram level. However, doing so for a non-stationary time series introduces the possibility of spurious variation of the

statistical distribution's parameters in Welch's method. Alternatively, Schulz (1997) applies variance normalization after Welch's method, which implies that for this technique, it is appropriate to compute statistical distribution parameters for the entire time series. Doing so avoids the potential pitfalls of calculating the parameters for individual time series segments.

Based on research of commonly used probability distribution functions, the 3-parameter gamma distribution appears to be a reasonable and robust choice. The standard 3-parameter gamma distribution is defined as

$$g(x) = \frac{e^{-\frac{x-\mu}{\vartheta}} (x-\mu)^{k-1}}{\Gamma(k) \vartheta^k}, \quad (2.47)$$

where μ , k , and ϑ are the location, shape, and scale parameters, respectively, and the ordinary gamma function is given as

$$\Gamma(k) = \begin{cases} (k-1)!, & k \in \mathbb{N}^+ \\ \int_0^\infty x^{k-1} e^{-x} dx, & k \in \mathbb{C}^+ \end{cases}. \quad (2.48)$$

The gamma cumulative distribution function $\dot{\xi}(x_n)$, illustrated in Figure 3, is defined as

$$\dot{\xi}(x_n) = \frac{1}{\Gamma(x_n - \mu)} \int_0^{\frac{x_n - \mu}{\vartheta}} x^{k-1} e^{-x} dx. \quad (2.49)$$

The integral in $\dot{\xi}(x_n)$ is the lower incomplete gamma function. If a closed form for the antiderivative of $\dot{\xi}(x)$ were available as, say, $\dot{\chi}(x)$, the rightmost term of Equation (2.46) would take the form

$$r_n \Theta([q_n - r_n]) (\dot{\chi}(x_n) - \dot{\chi}(\{x | q(x) = r_n\})). \quad (2.50)$$

There is no closed form for the quantile function of 3-parameter gamma distributions, and the additional complexity of the rainfall distributions $X_n(t)$ makes it prudent to use a numerical method to find the correct x_n for a given value of $Q_n(t) = F_n^{-1}(t)$.

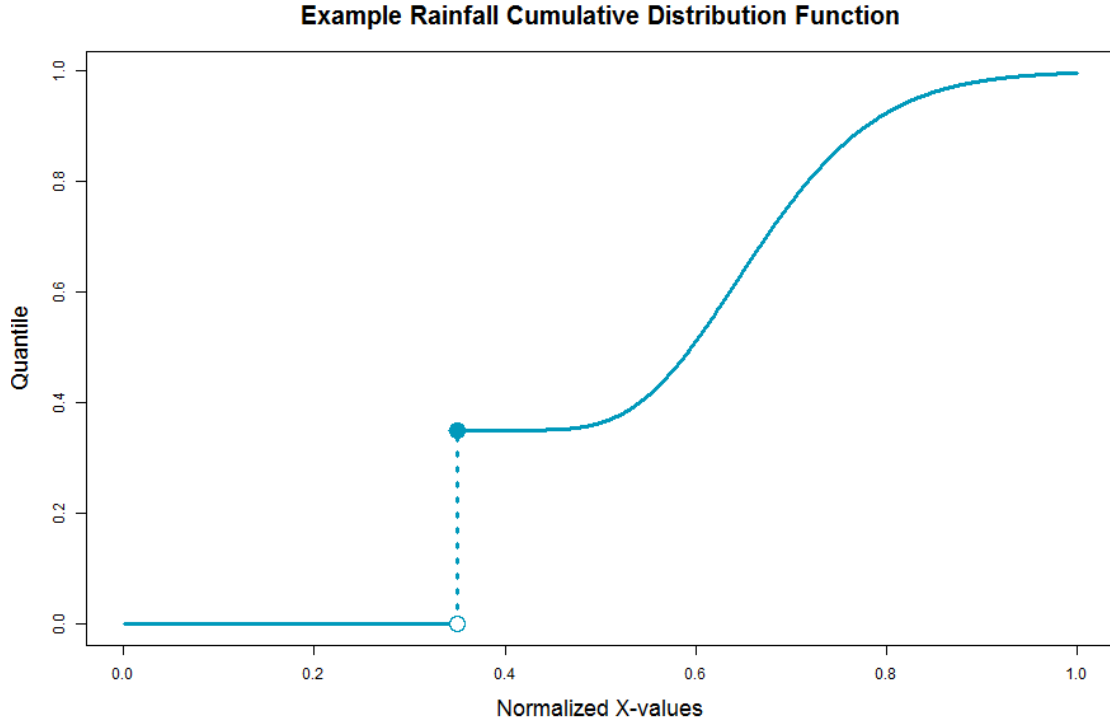


Figure 3: The general shape of a rainfall cumulative distribution, not including the autoregressive term. Since dry observations all have the same rainfall value (i.e., zero), the discontinuity represents the integration of a delta function in the probability distribution function.

Less exactly, x_n could also be estimated with an empirical cumulative distribution function (ECDF), thereby avoiding the conventional pitfalls of choosing a statistical distribution altogether. However, this is unlikely to yield realistically distributed values at the upper tail, since the ECDF will be sparsely defined there.

By the Central Limit Theorem though, it may be possible to identify one particular rainfall periodicity that accounts for most of the “gammaness” of rainfall variability, and by subtracting that gamma-distributed component from the raw data, it may be possible to model x_n as a normal distribution instead. Since this method is computationally the

simplest, I will choose this method if the rainfall data can be altered to reasonably approximate a normal distribution.

2.4 Computation and Benchmarking

Since the TRMM 3B42 data set is large (~500 Gb), I chose to keep the 3-hourly files in a compressed form, and create $10^\circ \times 10^\circ$ blocks of time series data one at a time. This reduces the disk space required, but it requires a decompression program, and since the decompressed files are iteratively deleted once the needed data is extracted, every pass requires that the files be decompressed again. While this method is not prohibitively time consuming, NCO operators (Zender, 2016) and similar utilities should be the preferred method when disk space is not a limiting concern.

As a consequence of programing LSDFT methods in both R and CUDA, it is possible to time the respective programs to measure their relative performance. I choose to use the default R compiler to enhance speed relative to uncompiled R while using only tools available to the typical user. Though the AR1 simulation method from SS02 is too computationally demanding to be useful for parallel computation in R, the basic LSDFT and WOSA (from SS97) methods can still be timed.

Computations have been performed on a machine with an AMD FX-9590 processor, which is clocked at 4.89 GHz and contains eight cores, and a GTX-670 GPU, which is clocked at 0.98 GHz and contains 1344 cores. Each block is parallelized in CUDA, but in R, each block is run on a single core parallel to the others. For comparability, I choose to only time the core time series analyses, rather than including data loading and other housekeeping functions. To compare the computing times meaningfully, I opt to examine the total timings

for each block of data, and scale these times according to clock speed and number of cores used. This ratio R can be expressed as

$$R = \frac{s_c n_c t_c}{s_g n_g t_g} = \frac{(4.89 \text{ GHz})(1) t_c}{(0.98 \text{ GHz})(1344) t_g}, \quad (2.51)$$

wherein s represents the clock speed, n is the number of cores, t is the time taken per block, and the subscripts c and g represent the CPU process (in this case, the R code) and GPU process (CUDA code) respectively.

3. RESULTS

In order to demonstrate the functionality of the time series algorithms coded for the present research, I choose to focus on the diurnal cycle, since it is one of the most prominent and predictable oscillations in atmospheric science.

An example WOSA spectrum is presented in Figure 4. The theoretical AR1 noise spectrum is demarcated with the solid blue line, while the 95% confidence interval is bounded by the dashed blue lines at ± 6 dB. Since values of τ are relatively small, the curvature of the AR1 noise distribution is not immediately noticeable, but does present at high frequencies. The diurnal cycle yields a prominent peak at the proper frequency.

The results for the classic LSDFT, WOSA, and AR1-reduced spectra are presented as maps in Figure 5, Figure 6, and Figure 7. Since subtracting the annual cycle appears to yield results sufficiently modeled with a normal distribution, it is possible to use a normally-distributed random number generator in the AR1 Monte Carlo simulation, as opposed to using the more complicated gamma-distributed random number generator. For comparability, I choose to demonstrate the LSDFT and WOSA methods on this deannualized data set as well.

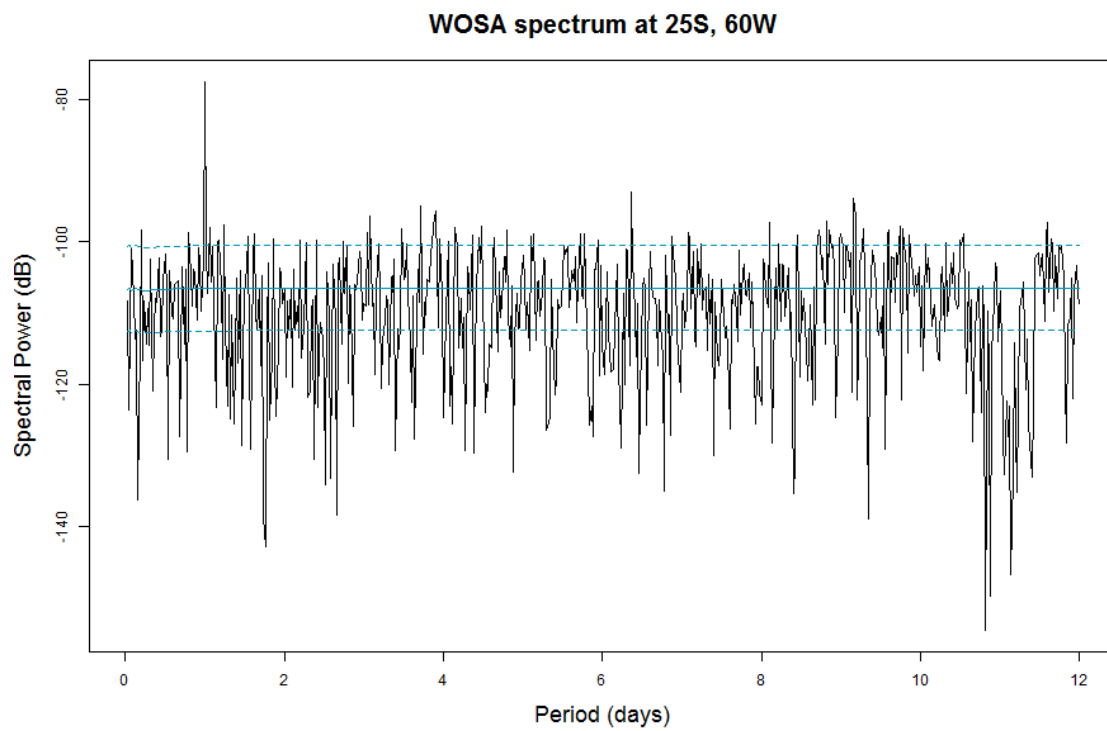


Figure 4: WOSA results with theoretical AR1 distribution as from SS02 in blue, and the two-sigma confidence limits as dashed lines.

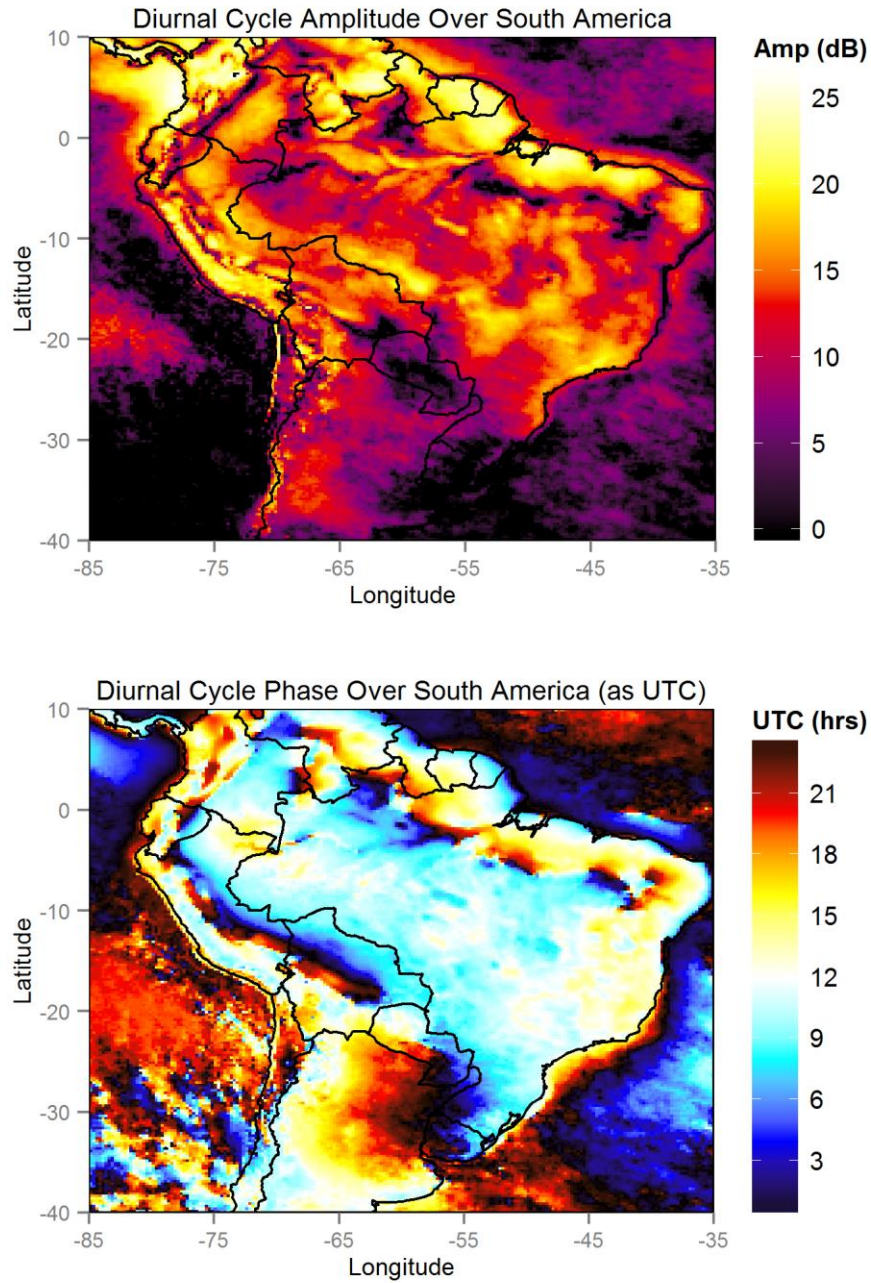


Figure 5: The classic LSDFT analysis with a complex formulation over South America. This analysis shows coherent spatial patterns at the frequency corresponding to the diurnal cycle in terms of both (top) relative amplitude, highlighting areas with strong diurnal cycles, and (bottom) phase, showing areas with propagating diurnal storm systems as rainbow patterns, such as over Argentina.

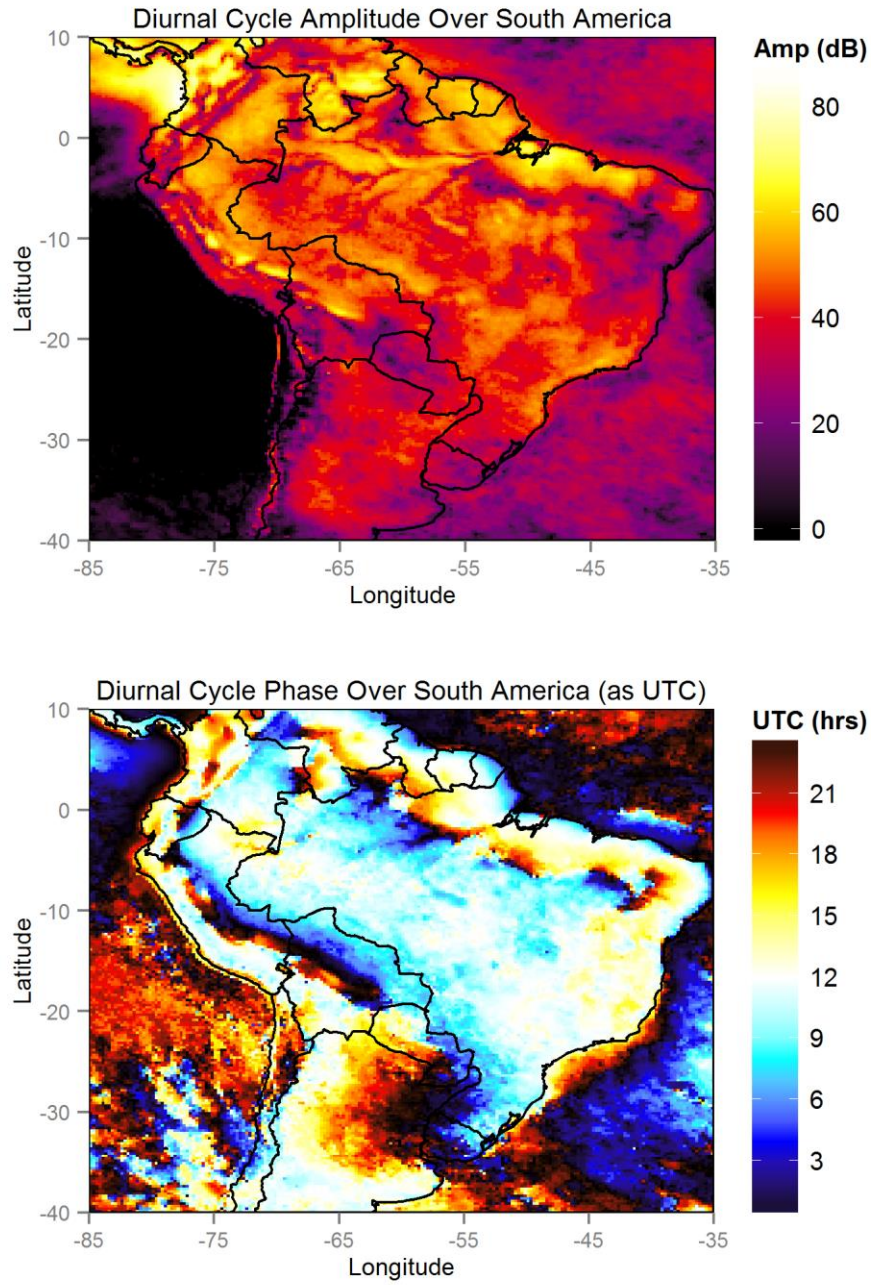


Figure 6: The WOSA method applied to the LSDFT (top) results in stronger relative amplitudes in most places, but appears to granulate spatial patterns in general, and (bottom) requires circular averaging of phases, which appears to granulate spatial patterns in phase as well.

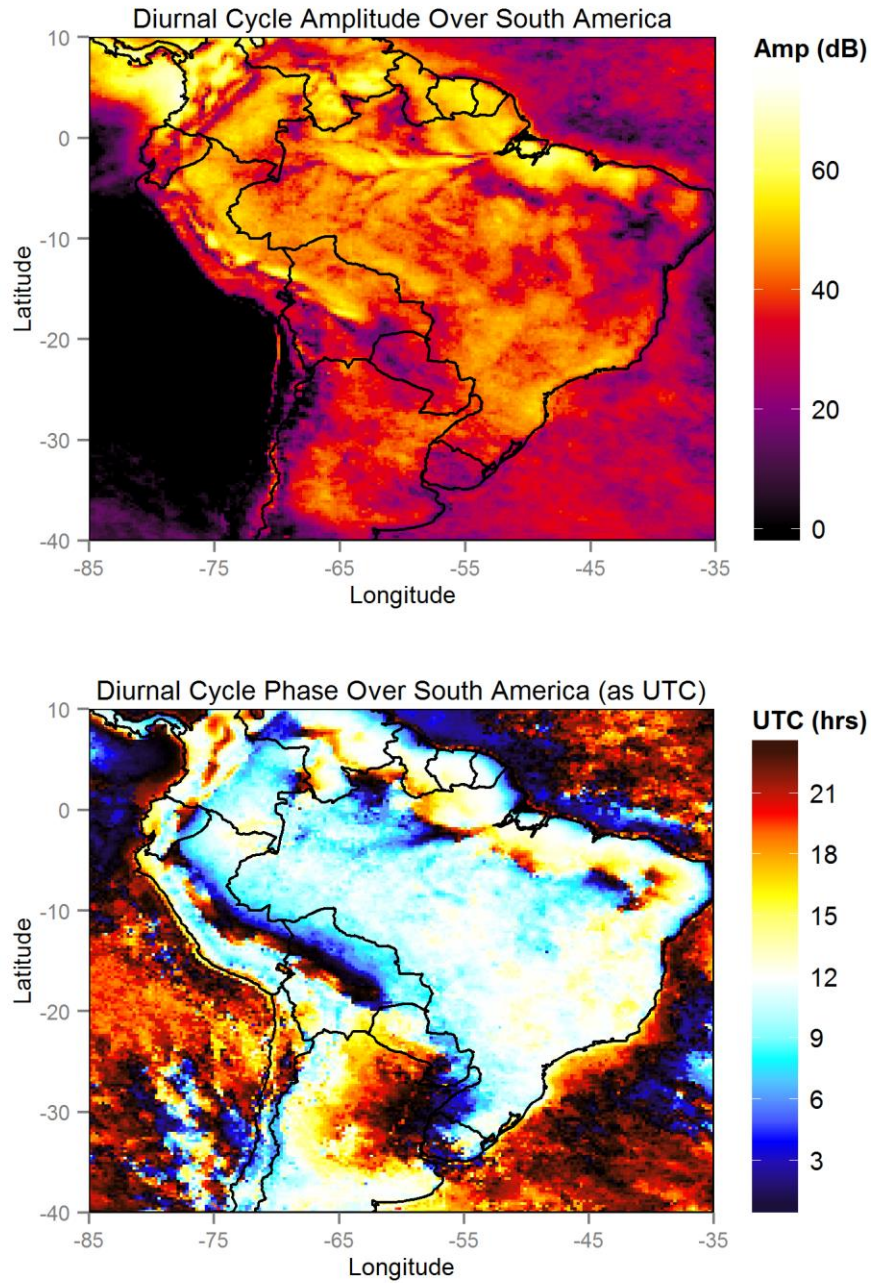


Figure 7: The WOSA method corrected for AR1 red noise estimates as per SS02. Correcting for AR1 according to the complex formulation introduced herein does not result in the same noise reductions that are typical for periodogram data. Simply borrowing phase data from WOSA analysis appears to be inadequate.

The EOF results for classic, WOSA, and AR1-reduced LSDFT time series analyses are presented in Figure 8, Figure 9, and Figure 10. Due to memory constraints in evaluating the EOF problem, the domain was reduced to an area covering most of the Amazon basin.

Integrating over the whole spectrum of interest can reveal more chaotic patterns that occupy a wide frequency band. However, unlike in POP analysis, the resulting EOFs do not yield information about the frequency bandwidths they are associated with. In this case though, it can safely be assumed that the diurnal cycle will be the strongest oscillation, and indeed, the first EOF in each analysis appear essentially identical in both amplitude and phase, and accounts for the vast majority of the variance contained in the zero to twelve day bandwidth. Due to negative eigenvalues, the final EOFs from classic LSDFT analysis also contain significant amounts of variance. However, only the final EOF (Figure 8) appears to have a spatially coherent pattern, which generally matches a diurnal pattern save a phase jump mid-domain. For comparison, the WOSA EOF analysis of only the diurnal cycle yields just one EOF of note, while the rest have explained variances many orders of magnitude smaller (Figure 11).

The next EOFs reveal distinct patterns within the zero to twelve day bandwidth that cannot be found in an EOF analysis of just the diurnal LSDFT results. A pattern that appears to be consistent with strong cold surges from further south (e.g., Lupo et al., 2001) corresponds to the second LSDFT EOF, the third WOSA EOF, and the fifth AR1-reduced EOF. The second EOF from WOSA analysis seems to be a harmonic of the diurnal cycle and is similar to the EOF pattern at the half-day frequency (Figure 12). This pattern is not found in the other analyses because WOSA analysis was performed with higher frequency resolution, and the semi-diurnal frequency was not evaluated. Higher-index EOFs appear to have multipolar forms that bear little relation to known real-world mechanisms.

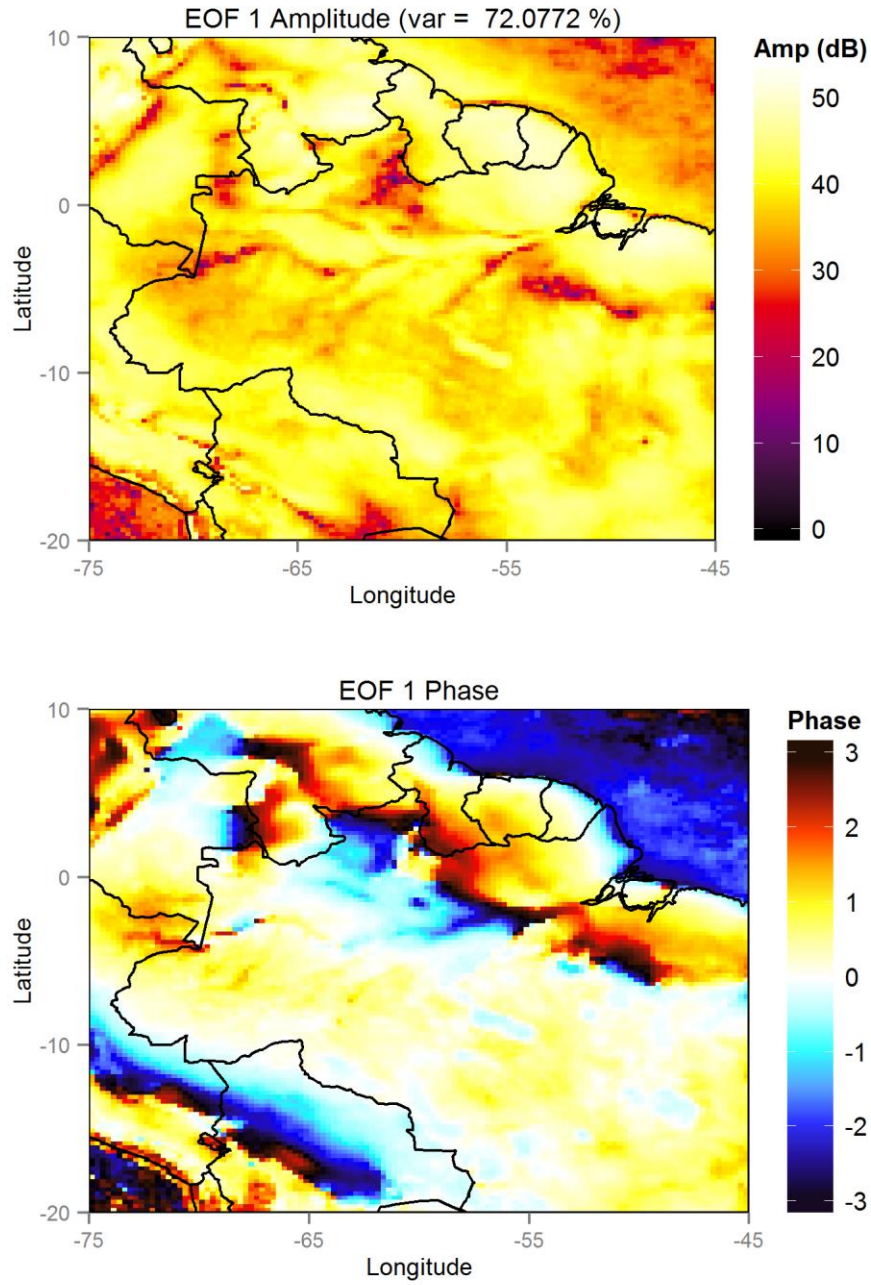


Figure 8: EOF results from classic LSDFT analysis, over the zero to twelve day bandwidth, showing relative amplitude and phase. Due to negative eigenvalues, some variance is apportioned to the last EOFs. The final EOF is closely matches the diurnal cycle, while the others are essentially noise profiles.

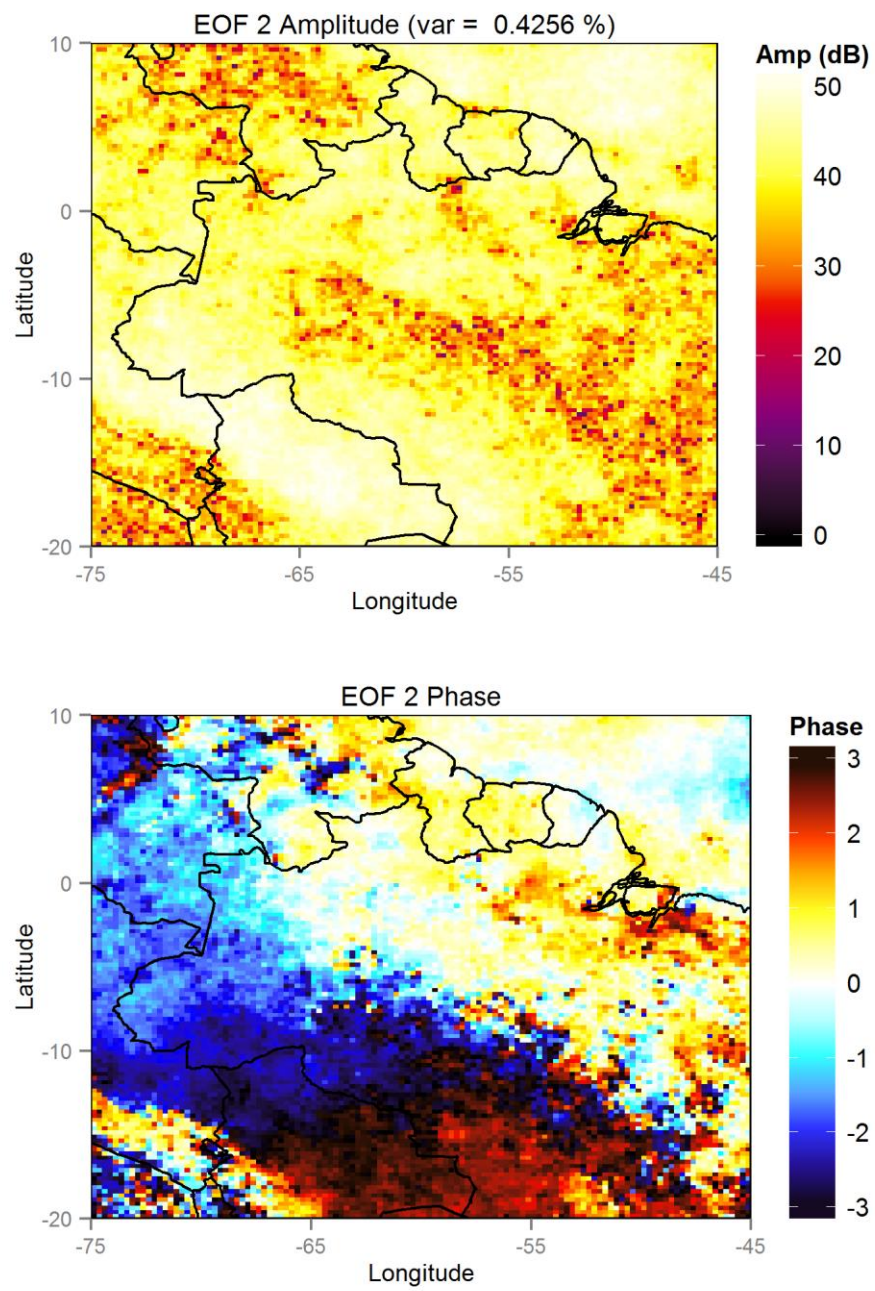


Figure 8: (continued)

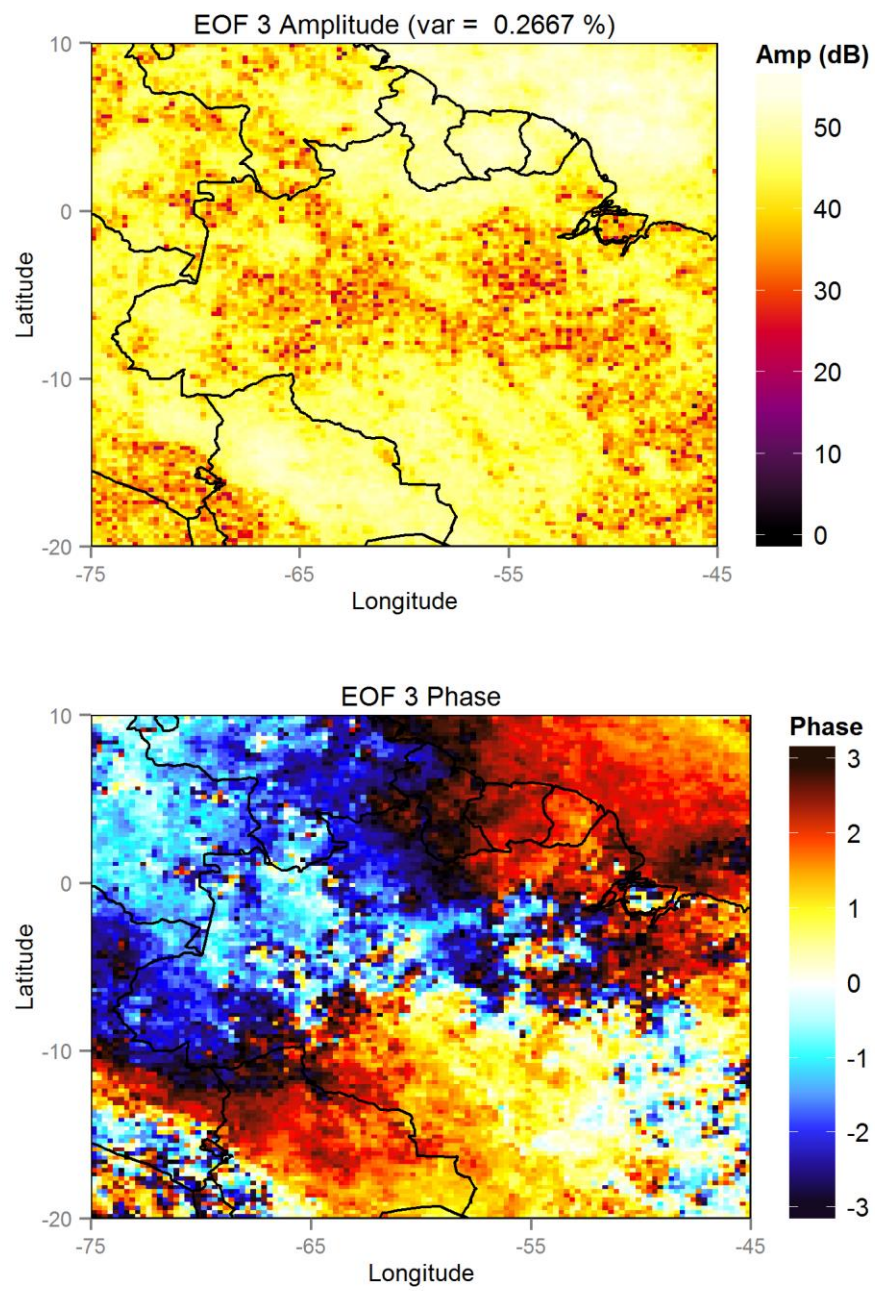


Figure 8: (continued)

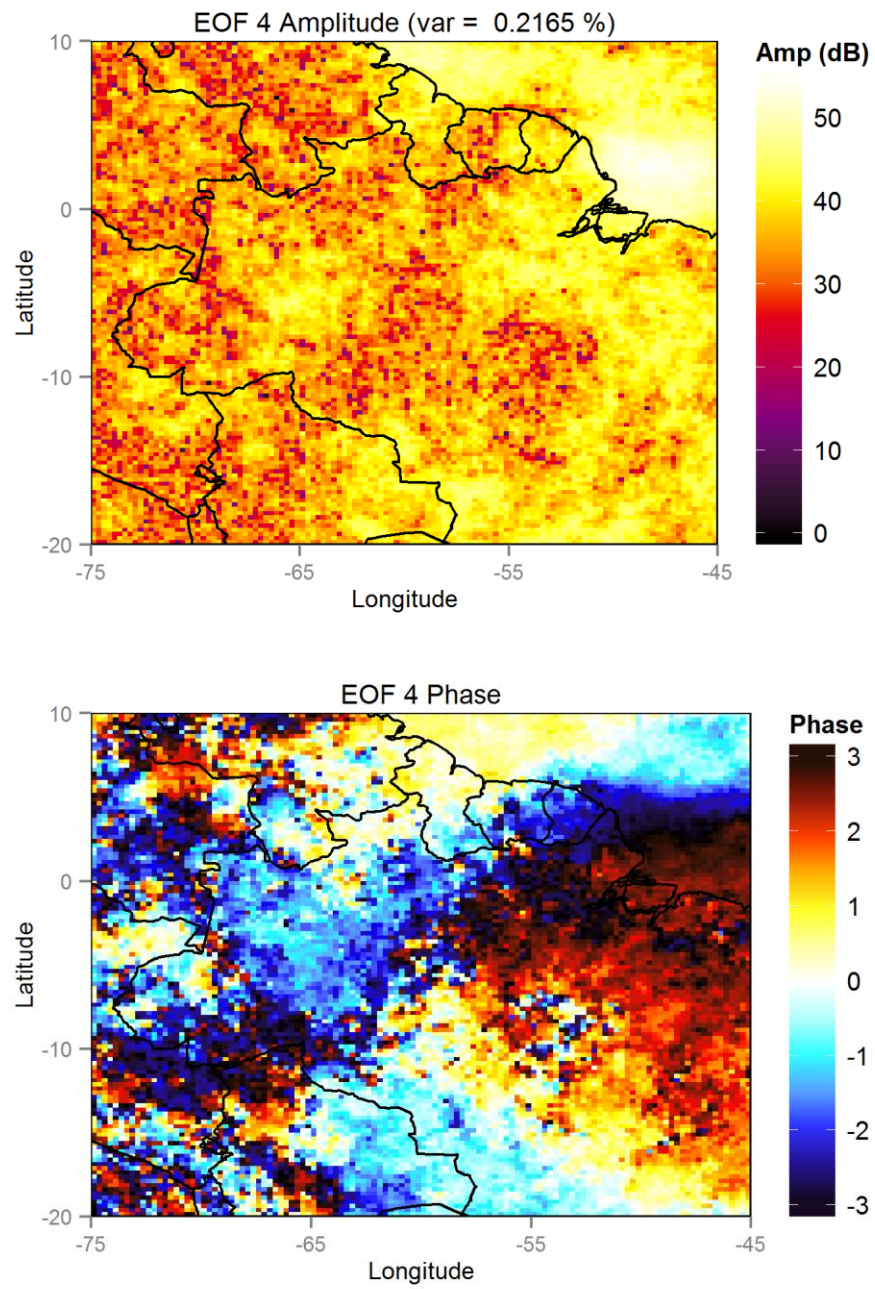


Figure 8: (continued)

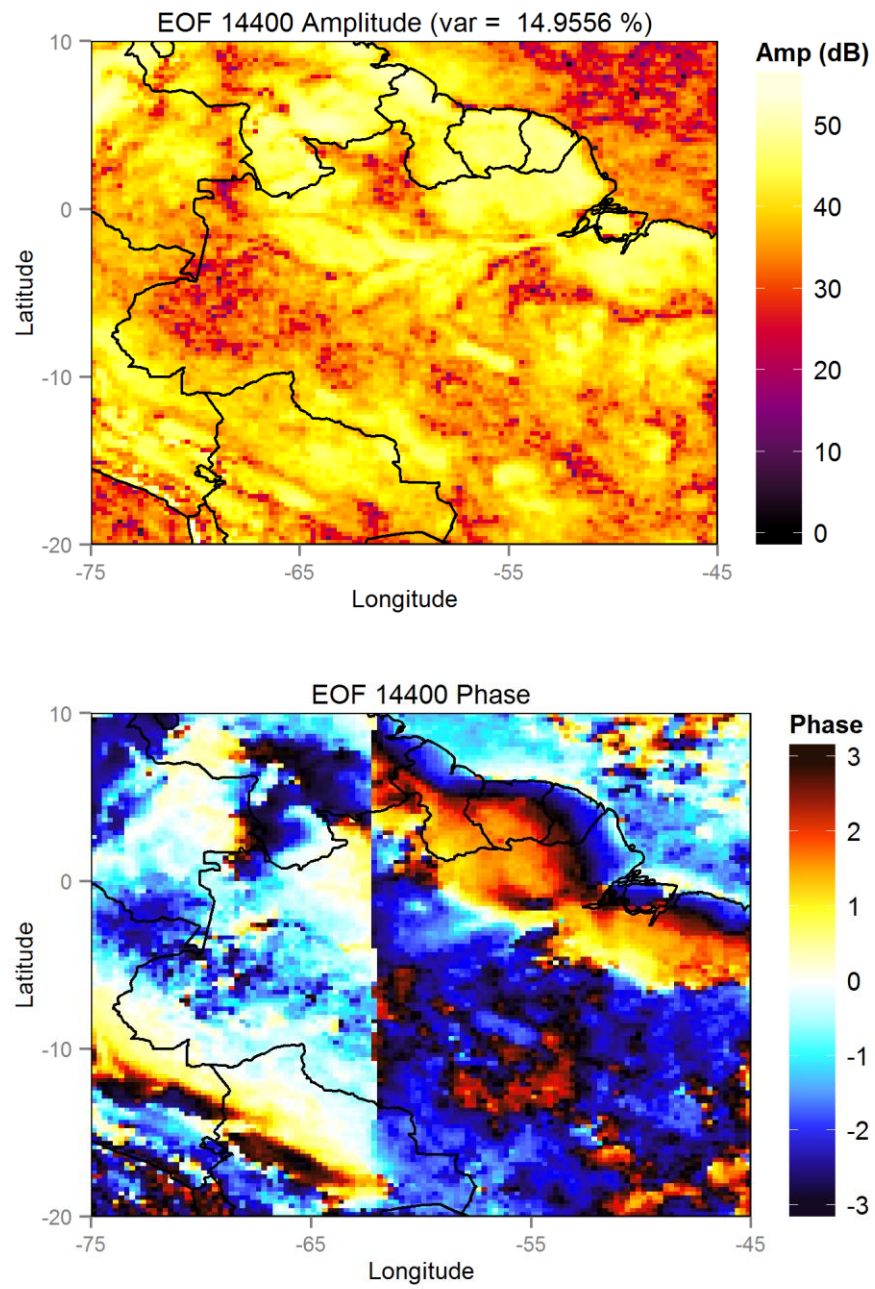


Figure 8: (continued)

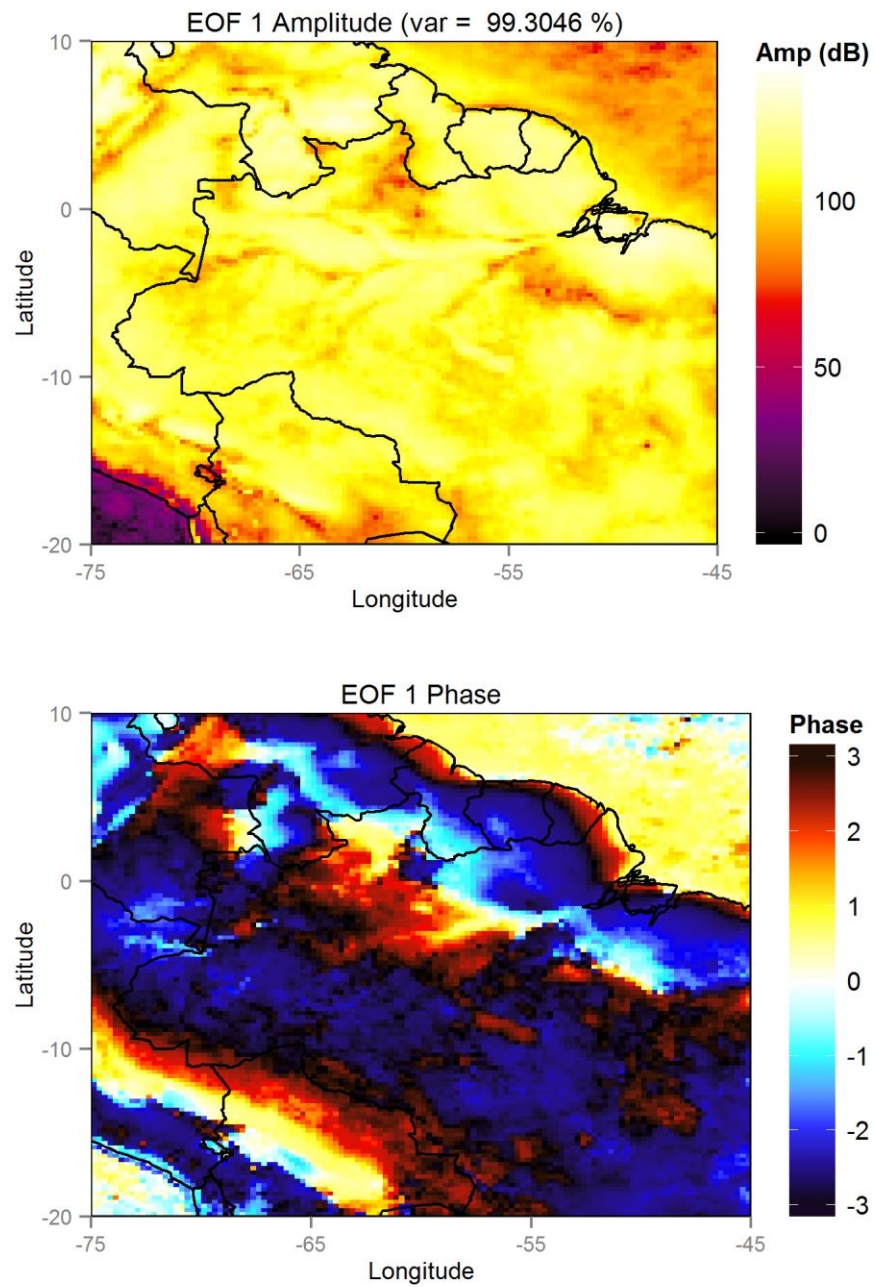


Figure 9: EOF results from WOSA analysis, showing relative amplitude and phase of the first four EOFs.

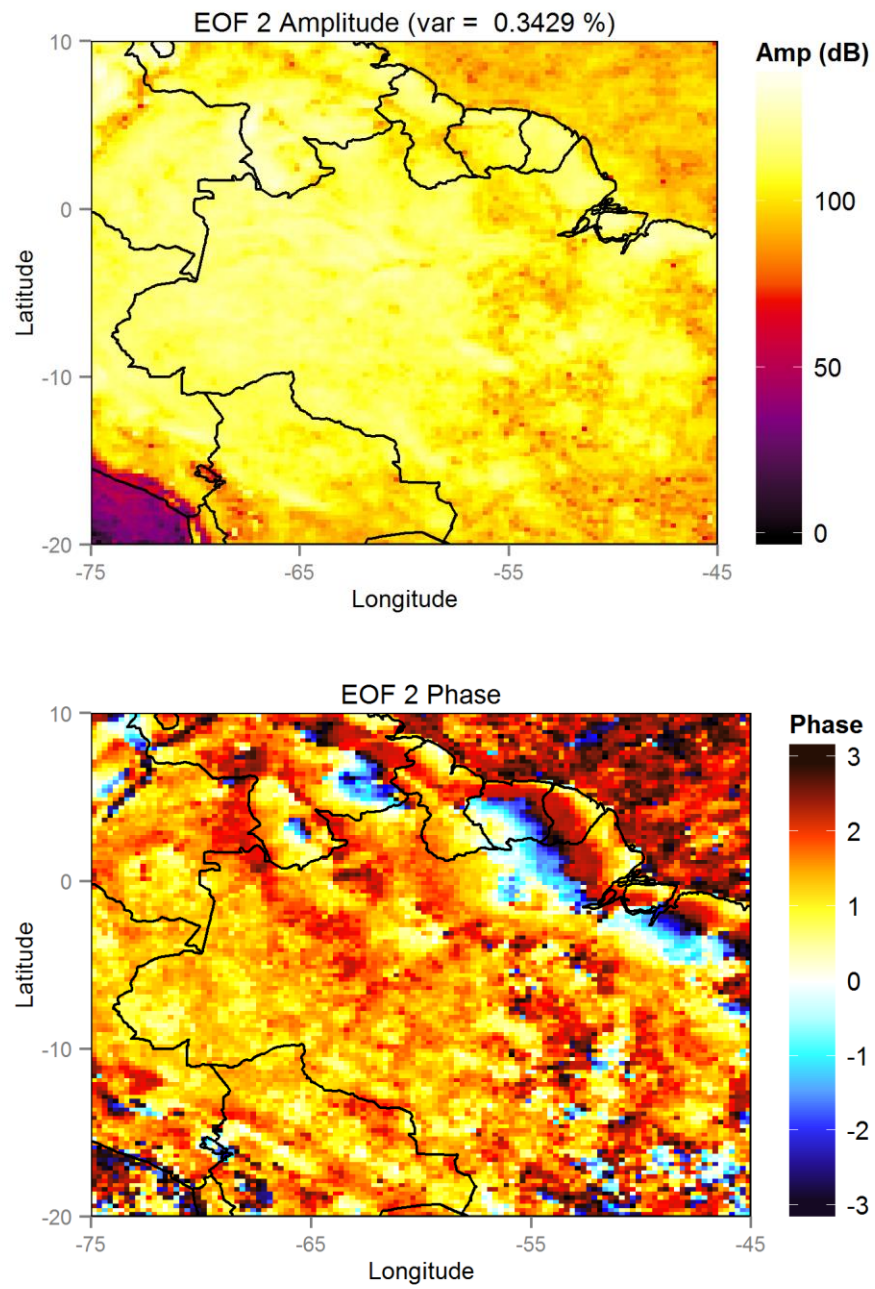


Figure 9: (continued)

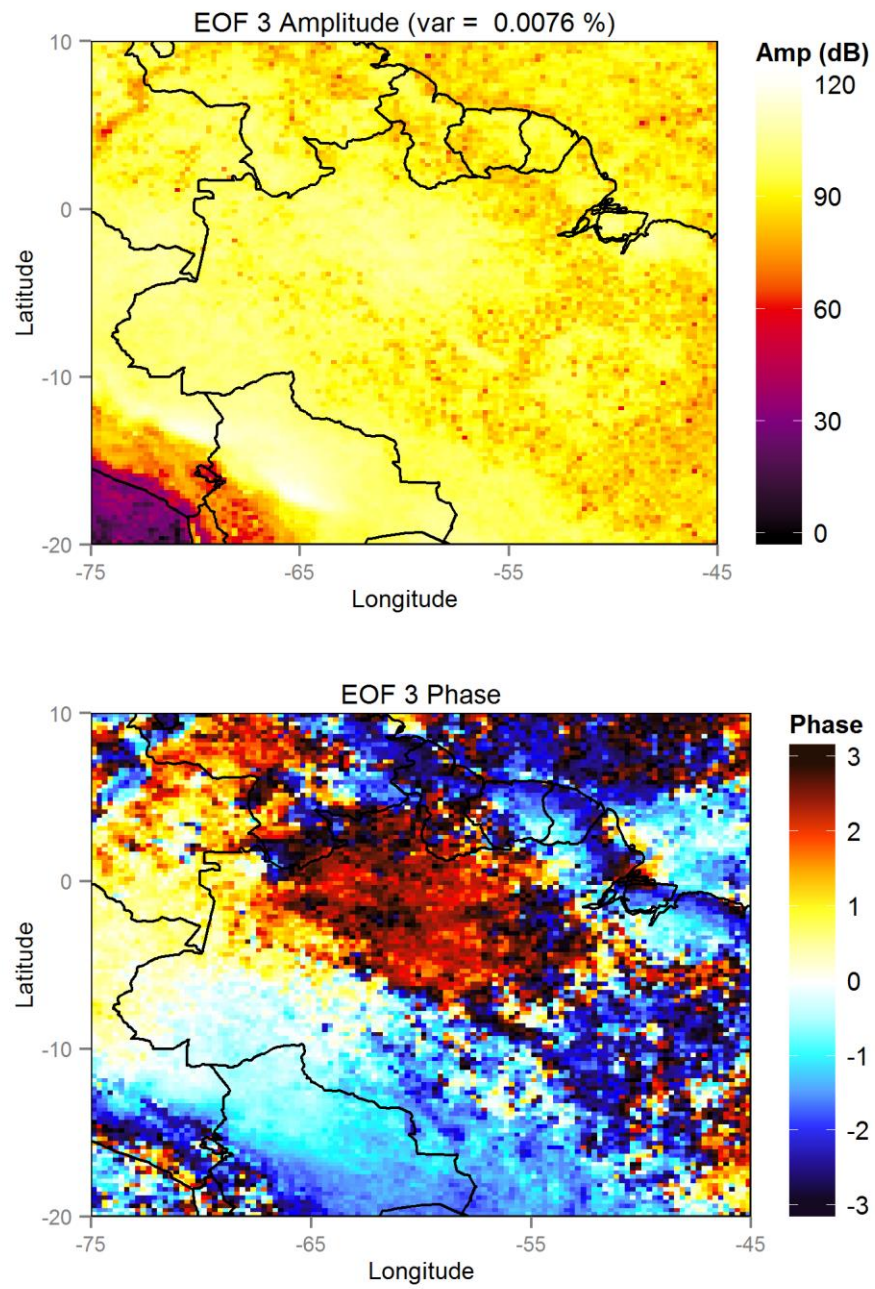


Figure 9: (continued)

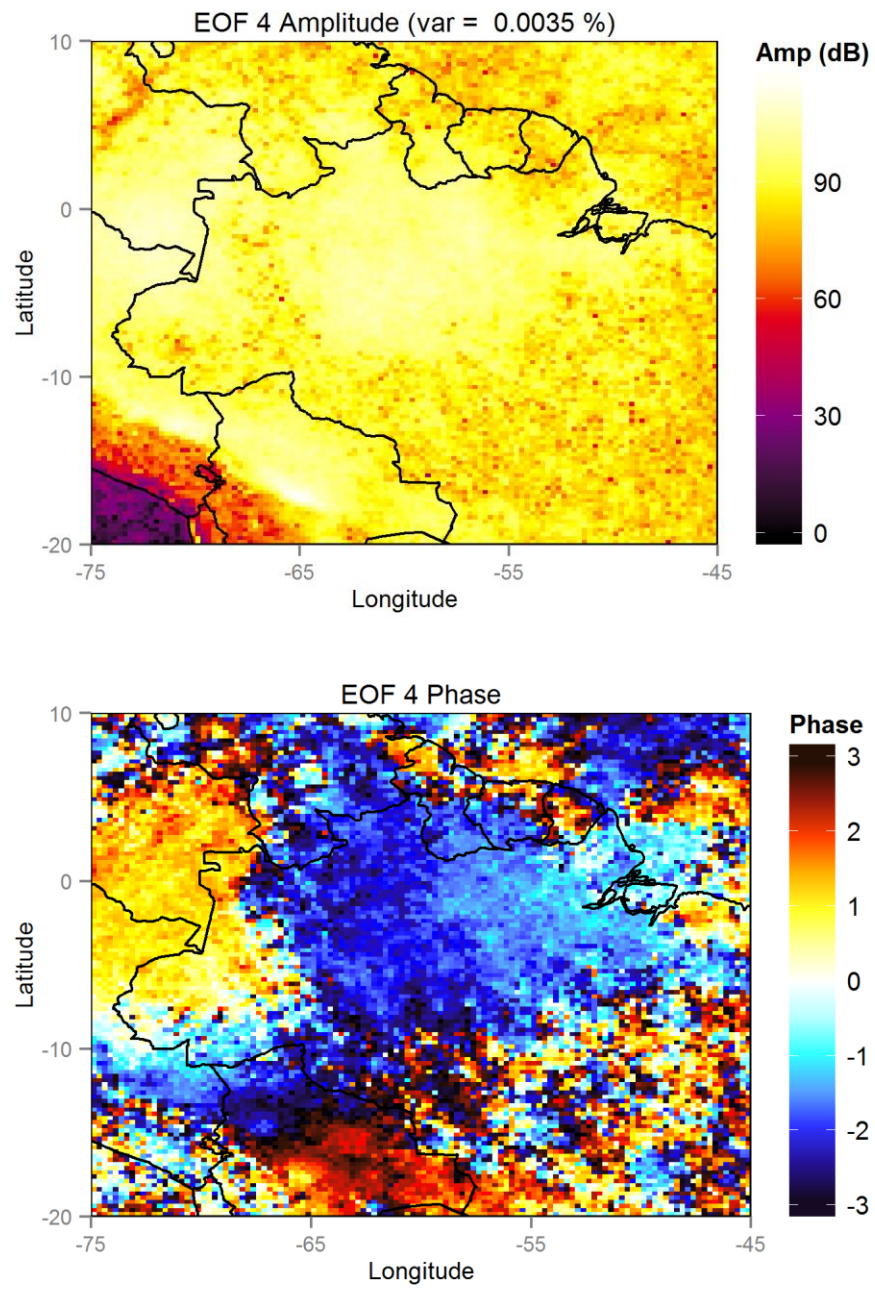


Figure 9: (continued)

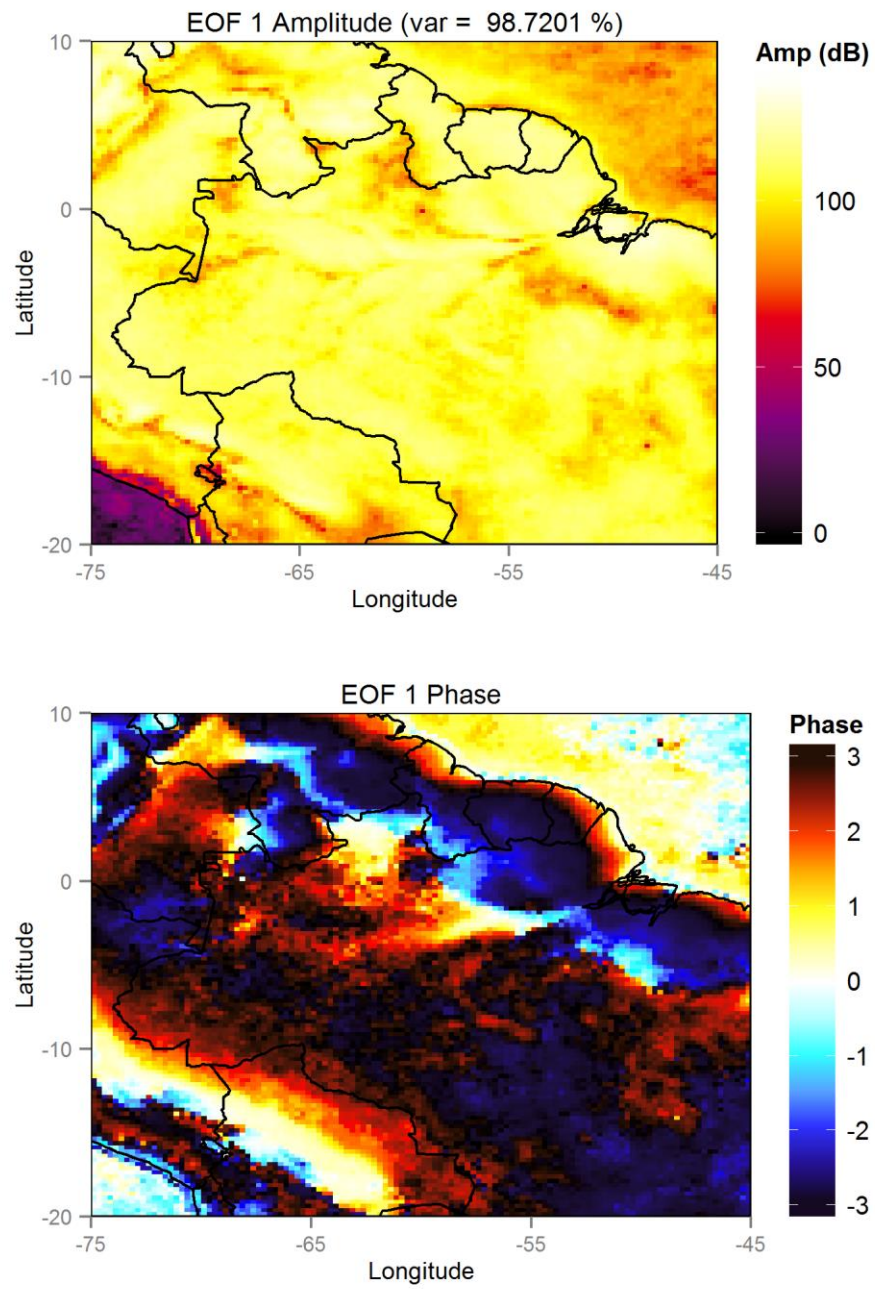


Figure 10: Selected EOFs resulting from WOSA analysis with AR1 reduction. EOFs 3 and 4 are essentially noise patterns.

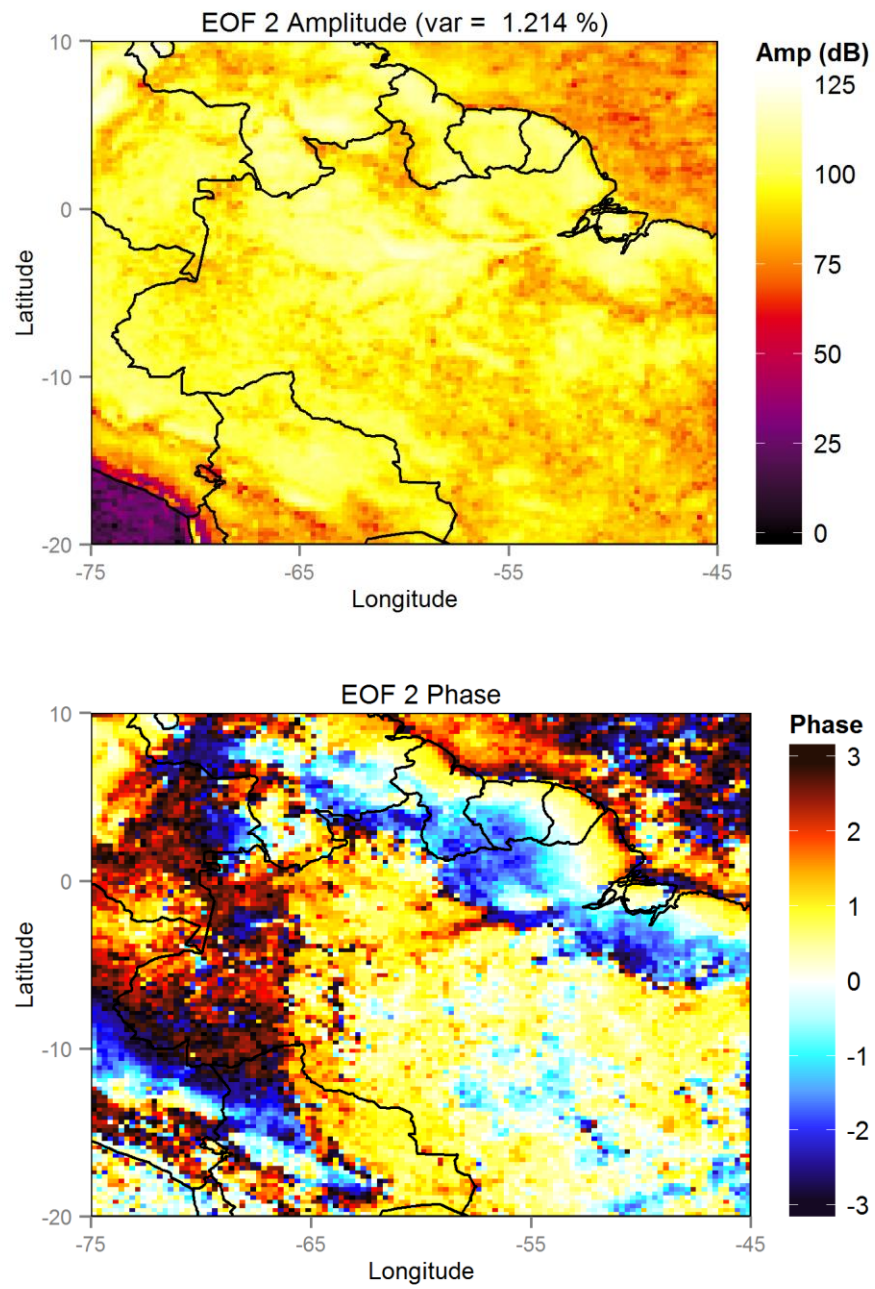


Figure 10: (continued)

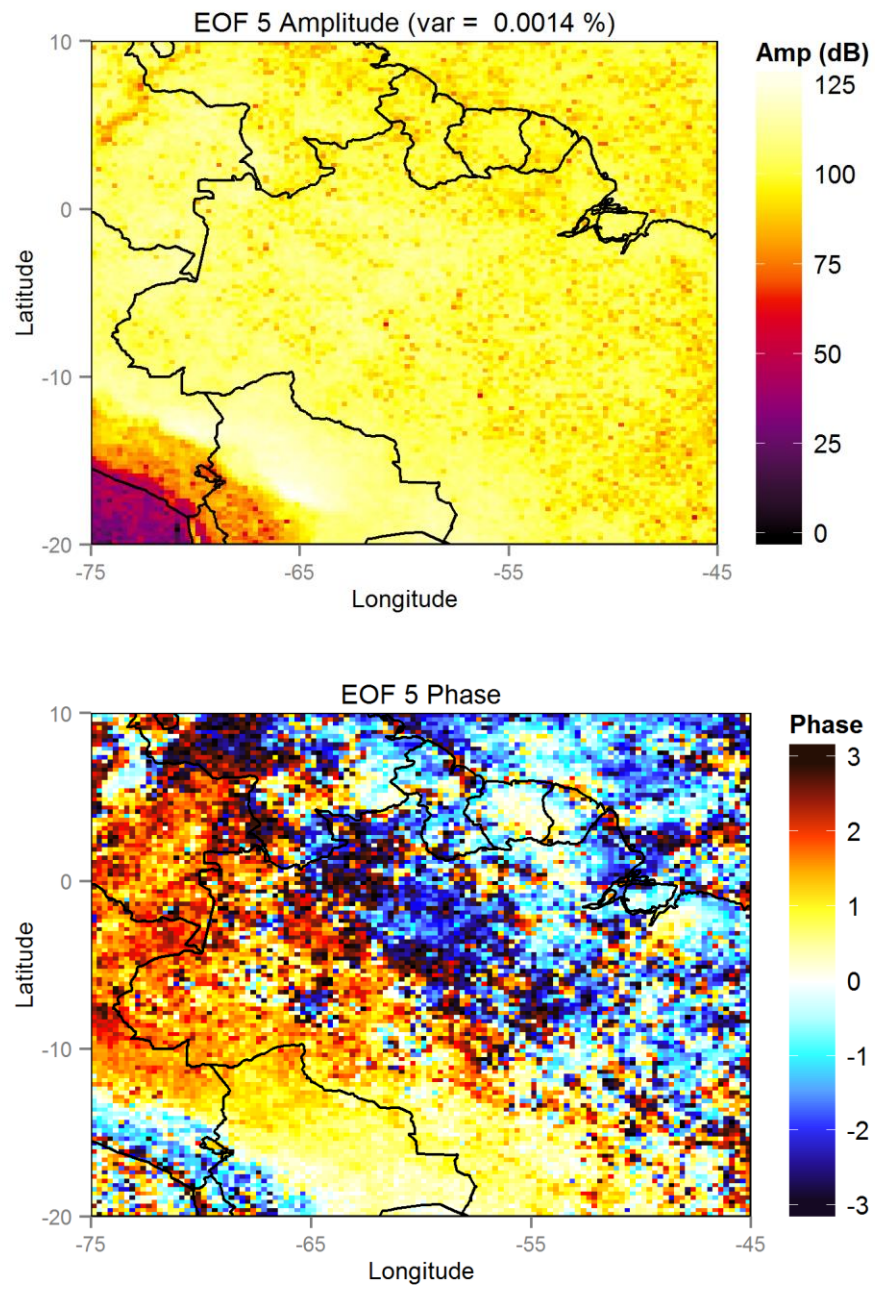


Figure 10: (continued)

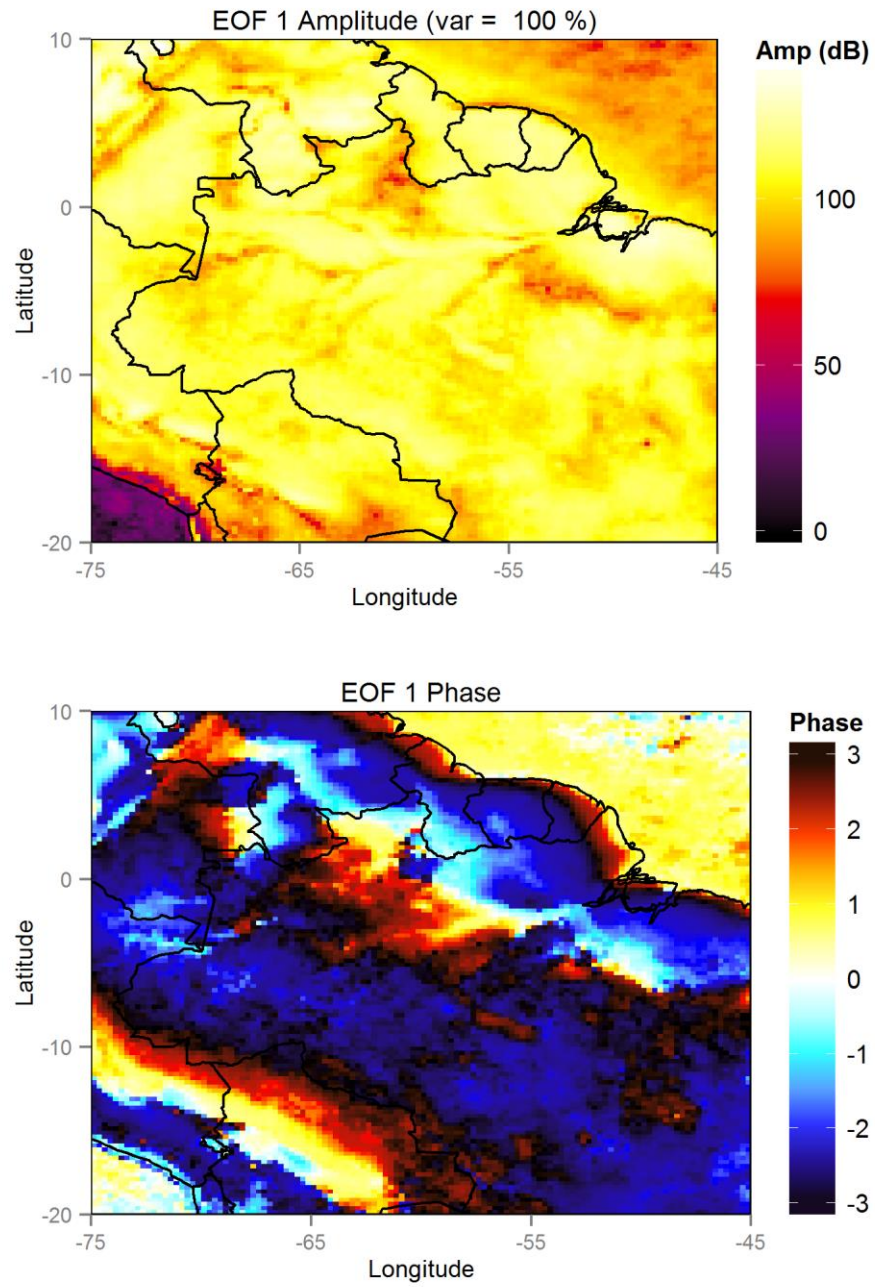


Figure 11: The WOSA EOF analysis of the diurnal cycle. The first EOF is the only notable EOF, as the next eigenvalue is fourteen orders of magnitude smaller than the first.

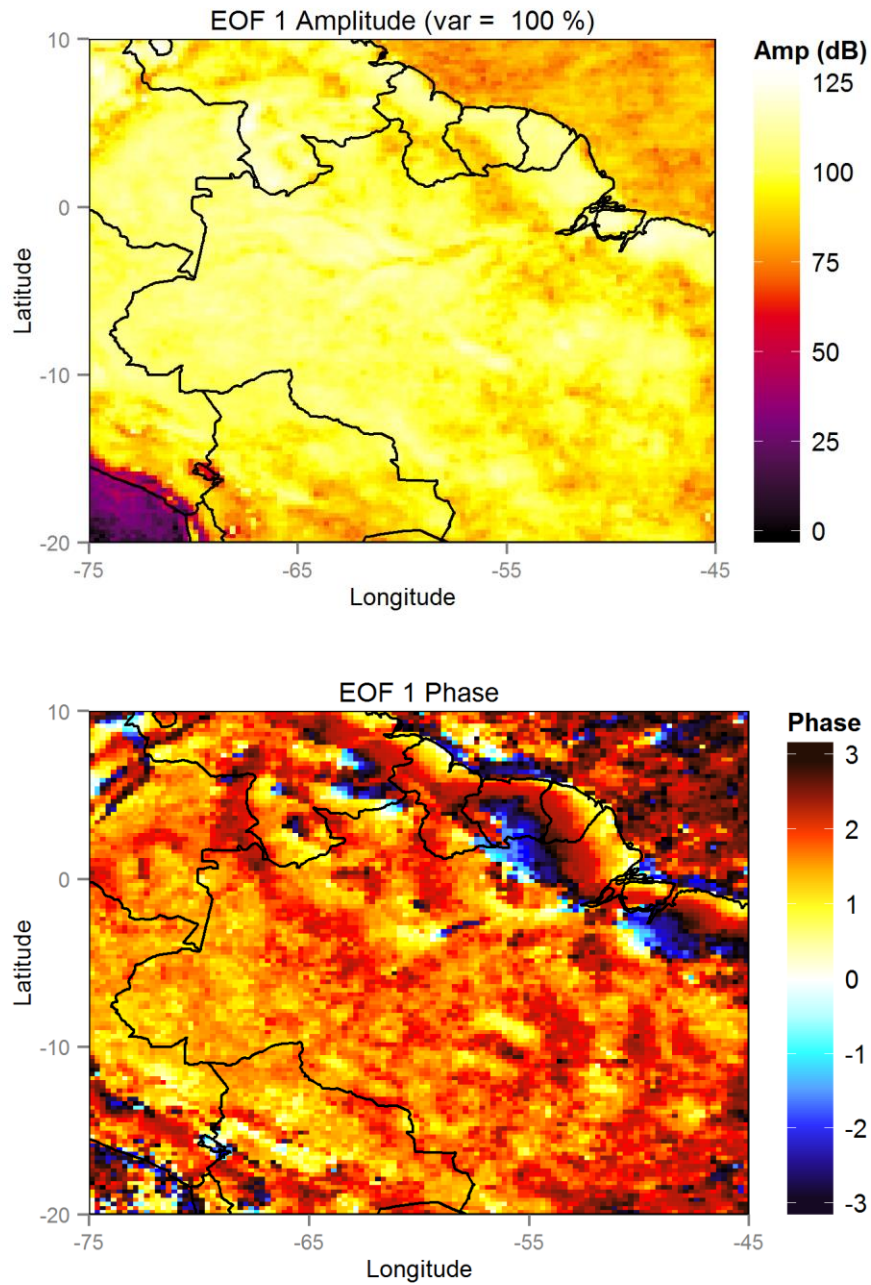


Figure 12: An EOF analysis of just the semi-diurnal component of WOSA analysis results in a pattern consistent with the second EOF of the integrated analysis from Figure 9. As in the EOF analysis of the diurnal frequency, this is the only significant EOF.

Because the data set contains a high number of observations, the error bars derived from North's Rule of Thumb (North et al., 1982) are difficult to resolve relative to the spread of eigenvalues. Therefore, I present the estimated confidence intervals for the three main analyses in tabular form in Tables 1, 2, and 3. These results show that interaction between the major EOFs should be minimal. WOSA EOF analysis appears to have the most rapid decay of variance with respect to EOF number, indicating that in the present research, WOSA offers the most minimally-descriptive EOF analysis.

The results of omitting a large portion of the time series appear to be basically similar to the results with no large gaps (Figure 13). This implies that there is no particular difficulty here as there can be with interpolation.

Table 1: Variances and Error Estimates for LSDFT EOF analysis

EOF number	Percent Variance	Variance Error
1	72.07716223 %	0.523750792 %
2	0.42558141 %	0.003092500 %
3	0.26674283 %	0.001938295 %
4	0.21647574 %	0.001573027 %
5	0.15662722 %	0.001138136 %
6	0.15054737 %	0.001093957 %

Table 2: Variances and Error Estimates for WOSA EOF analysis

EOF number	Percent Variance	Variance Error
1	99.998444671 %	0.7266416 %
2	$1.191973 * 10^{-3}$ %	$8.661508 * 10^{-6}$ %
3	$5.829745 * 10^{-7}$ %	$4.236201 * 10^{-9}$ %
4	$1.261664 * 10^{-7}$ %	$9.167920 * 10^{-10}$ %
5	$7.197626 * 10^{-8}$ %	$5.230176 * 10^{-10}$ %
6	$4.160337 * 10^{-8}$ %	$3.023121 * 10^{-10}$ %

Table 3: Variances and Error Estimates for AR1-reduced EOF analysis

EOF number	Percent Variance	Variance Error
1	98.72007 %	$2.583633 \times 10^{-3} \%$
2	1.214036 %	$2.865128 \times 10^{-4} \%$
3	$3.964074 \times 10^{-2} \%$	$5.177249 \times 10^{-5} \%$
4	$3.656714 \times 10^{-3} \%$	$1.572438 \times 10^{-5} \%$
5	$1.444182 \times 10^{-3} \%$	$9.881869 \times 10^{-6} \%$
6	$7.917555 \times 10^{-4} \%$	$7.316842 \times 10^{-6} \%$

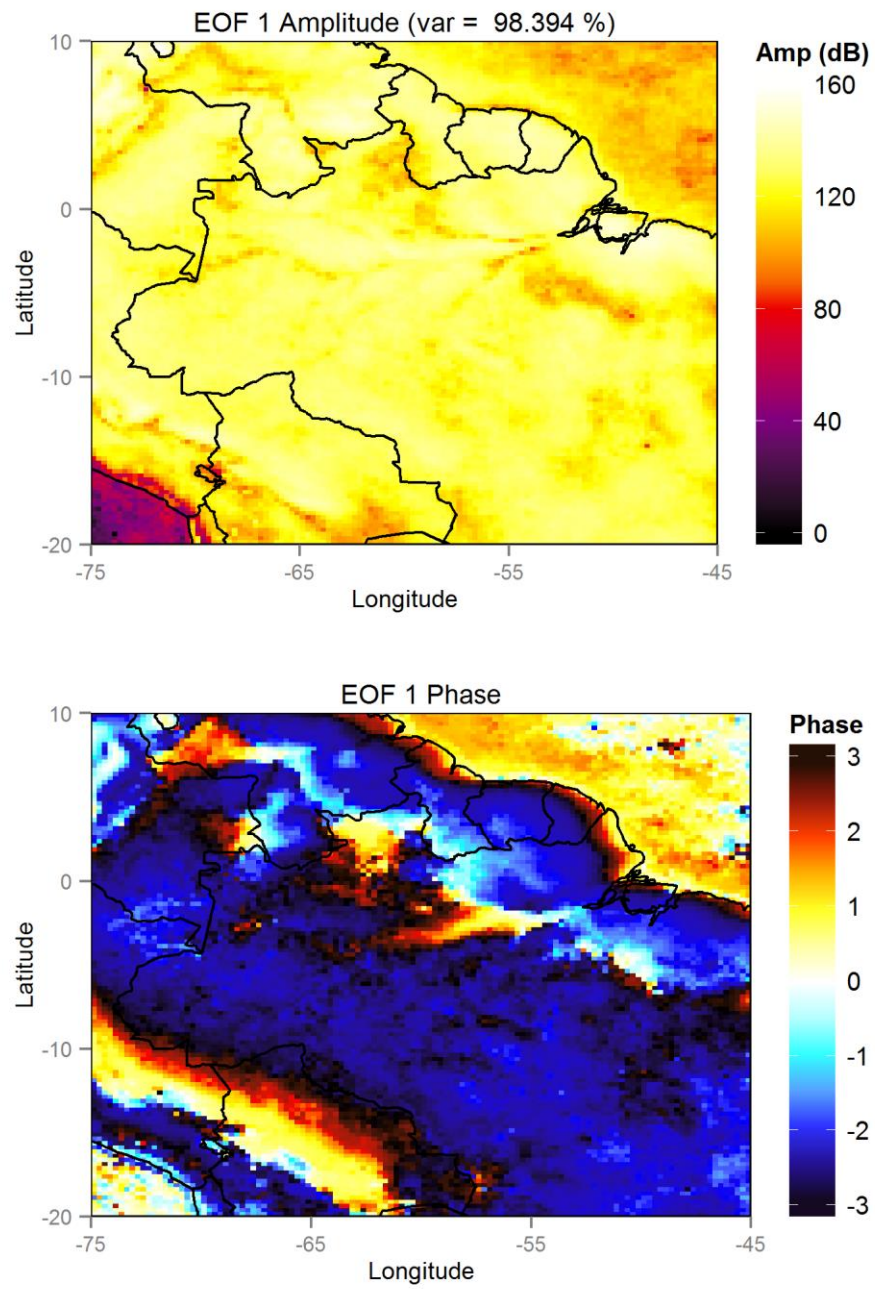


Figure 13: EOF results from WOSA analysis on the TRMM data with a large, artificial gap.

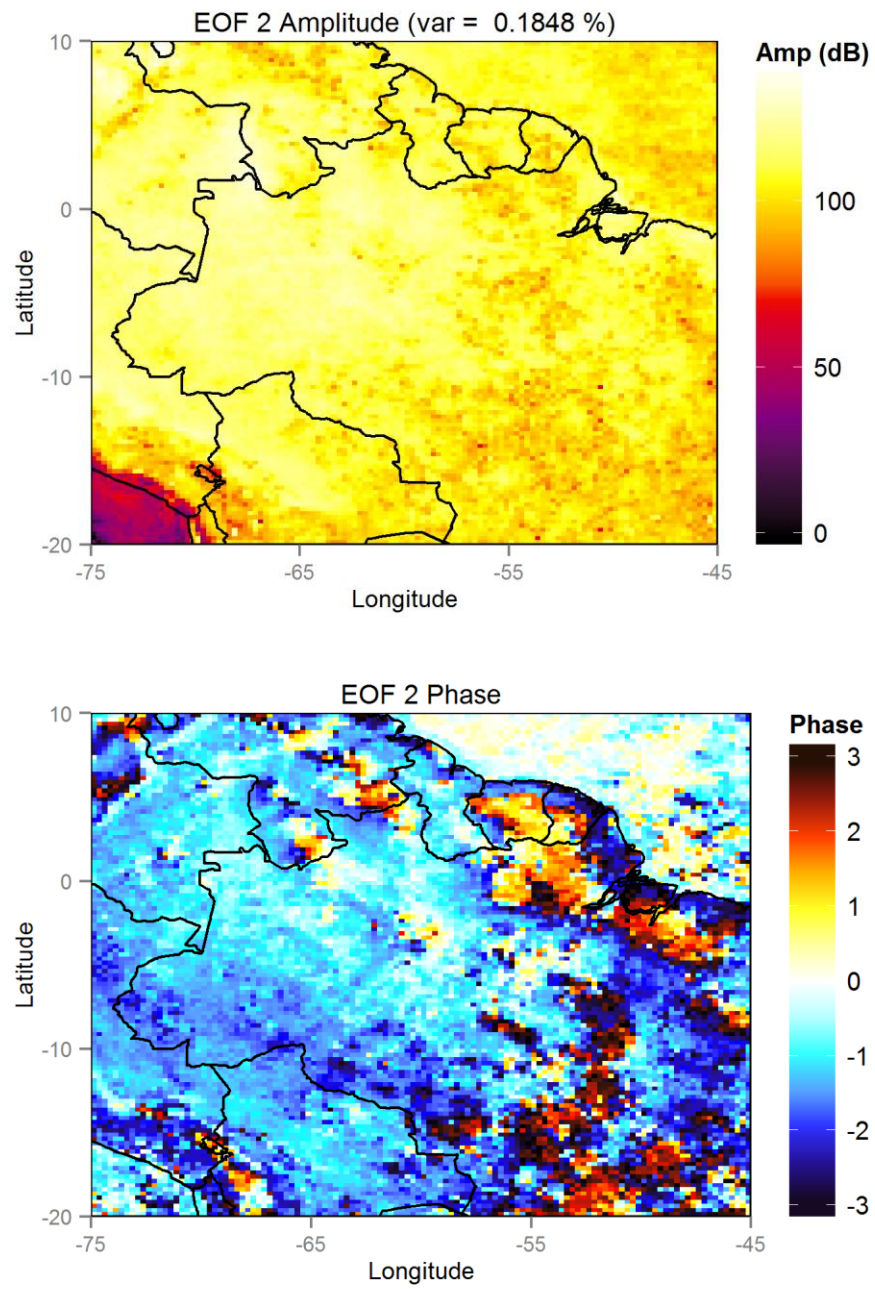


Figure 13: (continued)

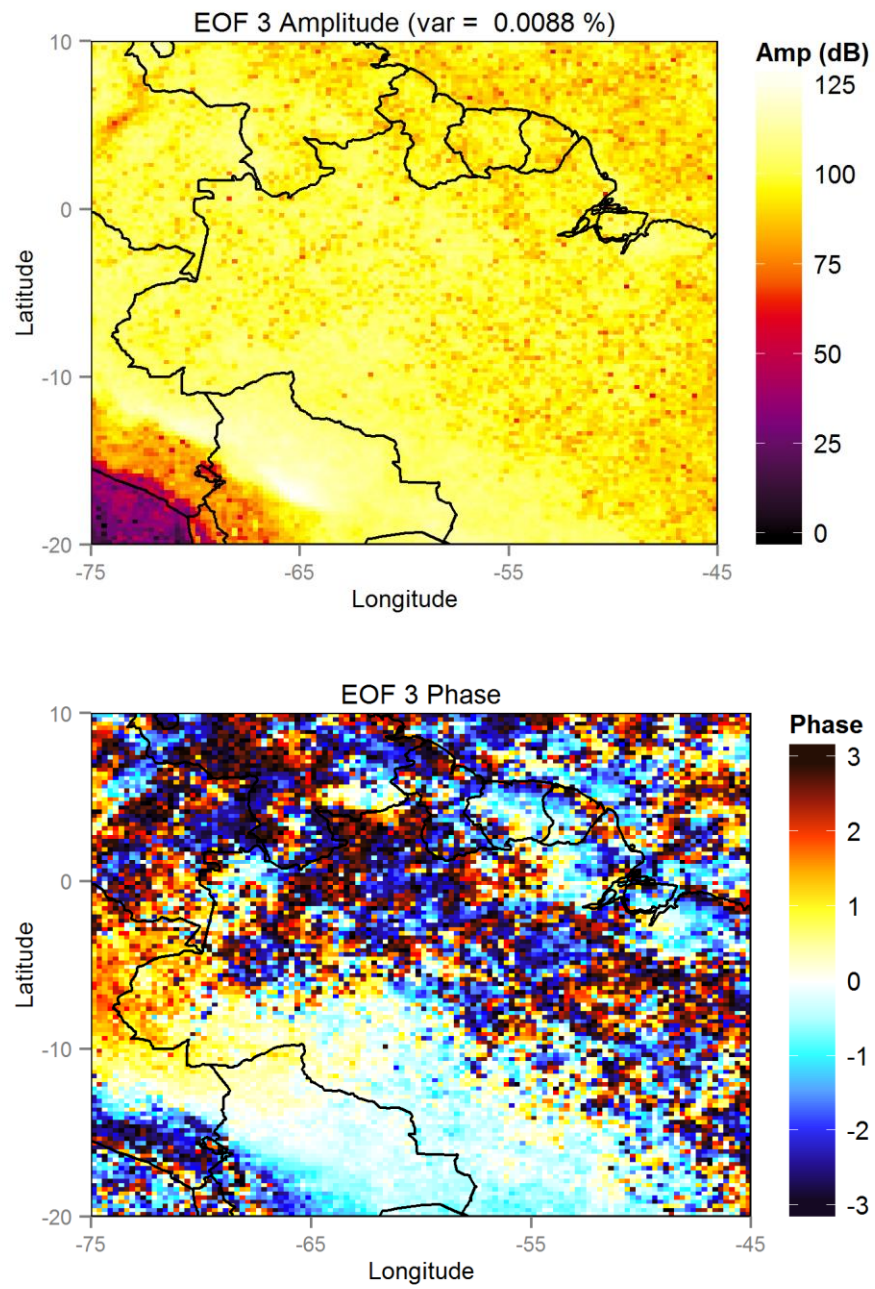


Figure 13: (continued)

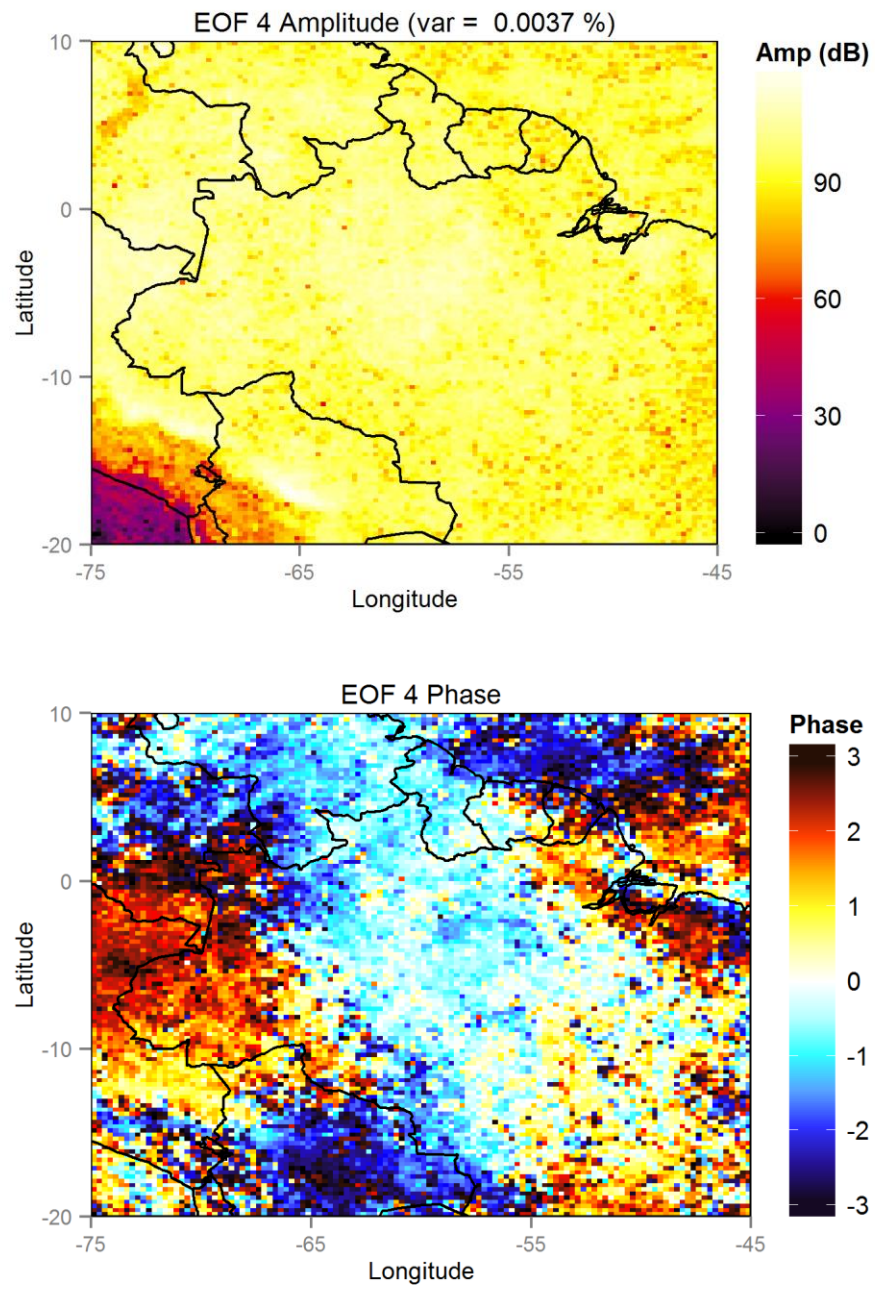


Figure 13: (continued)

3.1 Computational Results

Computational timings for each $10^\circ \times 10^\circ$ block in both R and CUDA are presented in Table 4. The speed ratio of CUDA to R, as defined in Equation (2.51), is also calculated for each block, averaging to 2.2 ± 0.1 .

Table 4: Timing of R and CUDA LSDFT algorithms

block	R time (s)	CUDA time (s)	Speed ratio
1-1	16761	27	2.3
1-2	15863	28	2.1
1-3	16651	29	2.1
1-4	16892	29	2.2
1-5	16927	29	2.2
2-1	16132	29	2.1
2-2	15289	28	2.0
2-3	16037	28	2.1
2-4	16420	28	2.2
2-5	16495	30	2.0
3-1	16653	29	2.1
3-2	15646	28	2.1
3-3	16527	30	2.0
3-4	16393	30	2.0
3-5	16890	30	2.1
4-1	17654	28	2.3
4-2	16800	28	2.2
4-3	17257	29	2.2
4-4	15995	30	2.0
4-5	17501	31	2.1
5-1	17770	25	2.6
5-2	17103	28	2.3
5-3	17581	29	2.3
5-4	17057	30	2.1
5-5	18033	30	2.2

4. DISCUSSION

As mentioned previously, EOF patterns obtained from the integrated covariance functions may obscure the relation of EOF patterns to frequency. A solution to this would be to solve the EOF problem for each frequency before integrating for comparison, but this can be computationally expensive for large frequency bands and high frequency resolutions. Additionally, it would only help in specifying strongly frequency-dependent patterns; more statistically chaotic patterns associated with a broad frequency band are liable to go undetected in the individual frequency-dependent analyses.

The inherently discrete nature of observational data implies that Fast Fourier Transform (FFT) algorithms and their derivatives will never be integrable across the entire range of frequencies. Therefore, it is important to note that the EOFs derived from integrating LSDFTs inherit the frequency window used in calculating the LSDFT, and cannot account for all possible variance in a time series set. However, this is only a minor concern as long as the EOFs of interest correspond to frequencies within the LSDFT frequency window. EOFs corresponding to frequencies outside of the frequency window will not be present, unless by aliasing.

The spatial patterns of the diurnal cycle are consistent with specific meteorological mechanisms (Yang and Slingo, 2001; Mapes et al., 2003). Notably, while a sea-breeze mechanism is evident over land near the coasts during the day, nocturnal off-shore propagation is also present in many areas. The strong amplitude off the west coast of Colombia is consistent with the results of Mapes et al. (2003), and the phase data derived herein presents a new way to view its propagation.

Using the classic LSDFT kernel is computationally expensive; indeed, that is the impetus for translating the LSDFT algorithms into CUDA. Faster algorithms do exist, notably the algorithm pioneered by Press and Rybicki (1989) and a number of similar extant techniques are collectively known as NFFTs. As these methods operate at the individual time series level, it is possible to replace the LSDFT kernel while still including WOSA and AR1 reduction techniques for greater accuracy. However, a key weakness of these methods is that for transforms of varying Nyquist frequencies, the frequency selection will be different. Therefore, it will be non-trivial to accurately compare the spectral power of multiple time series at a set frequency. However inconvenient, it should not significantly impact any resulting EOF analysis if the spacing is relatively consistent (i.e., as opposed to having large gaps relative to the temporal observation window). There may be a way to parameterize a time series' gappiness in terms of large gaps and spacing consistency, and a way to relate that to the comparability of NFFT segments, but that is not generally relevant to the basic algorithm described herein.

Computing the covariance matrix, either by direct calculation or by some other method, can be very memory-intensive. To reduce the memory overhead, it is possible to compute covariance elements in block form; that is, by only importing the time series data relevant to the area of the covariance matrix the program is iterating through. This requires opening either one or data blocks at a time, rather than requiring the outer product of the entire data set to be allocated at once. As above, the trade-off is that this block calculation of the covariance matrix can be time-intensive, so if method is necessary, care should be taken to make the blocks large enough that this technique does not result in a high number of block iterations.

Although the above method provides a path to large covariance matrices, the whole covariance matrix may be so large as to be impossible to load at once. The classic eigenvalue solvers usually require the entire covariance matrix to be loaded at once, but iterative solvers may be able to accommodate blocked data.

Choosing to use the univariate integral in LSEOF analysis may complicate and muddle the phase data unnecessarily compared to the bivariate integral, however, since the bivariate integral is very computationally expensive, a spatial comparison of TRMM data processed through both methods remains unfeasible at the present.

When the core LSDFT, WOSA, and AR1 routines are properly compartmentalized, it is possible to extend this software package as a whole to other time series problems. For example, Keplerian periodogram analysis is possible by replacing the LSDFT function (Zechmeister and Kürster, 2009), but the WOSA routine can still be used for noise reduction.

The LSEOF algorithm is closely related, but not identical, to the complex Hilbert EOF (HEOF) method. Since the HEOF method relies on a Hilbert transform of the raw data (Horel, 1984; Kim and Wu, 1999), followed by a classic EOF matrix multiplication, it arrives at the EOF covariance matrix in a different way from the LSEOF. Nevertheless, the Hilbert transform is a multiplier operator of the Fourier transform; that is,

$$F(H(u)) = -i \Theta(\omega) F(u), \quad (4.1)$$

where $H(u)$ is the Hilbert transform of the function u , $F(u)$ represents the Fourier transform of u , and Θ represents the Heaviside step function as classically defined, or more simply, the sign of its operand. The complex EOF results from adding real-valued data to its Hilbert transform, and treating the result of the data matrix of interest (Horel, 1984; Kim and Wu, 1999):

$$u' = u + H(u), \quad (4.2)$$

so

$$F(u') = F(u) + F(H(u)) = (1 - i\Theta(\omega))F(u) \quad (4.3)$$

and HEOF covariance matrix element σ'_{xy} is therefore

$$\begin{aligned} \sigma'_{xy}(t_{lag}) &= \int_0^\infty \tilde{X}(\omega) \tilde{Y}^*(\omega) e^{i\omega t_{lag}} d\omega \\ &= \int_0^\infty (1 - i\Theta(\omega)) \tilde{X}(\omega) (1 + i\Theta(\omega)) \tilde{Y}^*(\omega) e^{i\omega t_{lag}} d\omega \\ &= \int_0^\infty (1 + \Theta^2(\omega)) \tilde{X}(\omega) \tilde{Y}^*(\omega) e^{i\omega t_{lag}} d\omega, \end{aligned} \quad (4.4)$$

and since $\Theta^2(\omega) = 1$ as long as $\omega \neq 0$,

$$\sigma'_{xy}(t_{lag}) \cong \int_{0^+}^\infty (1 + \Theta^2(\omega)) \tilde{X}(\omega) \tilde{Y}^*(\omega) e^{i\omega t_{lag}} d\omega = 2 \int_{0^+}^\infty \tilde{X}(\omega) \tilde{Y}^*(\omega) e^{i\omega t_{lag}} d\omega \quad (4.5)$$

which implies that

$$\sigma'_{xy}(t_{lag}) \cong 2\sigma_{xy}(t_{lag}). \quad (4.6)$$

Since the covariance matrices are identical aside from a scalar multiple, the eigenvalues and eigenvectors of complex EOF analysis are identical to the results from LSEOF analysis.

The present research only makes use of the most basic EOF analysis, so the effects of using LSDFT analysis together with more advanced EOF/PCA analyses remain unknown.

5. CONCLUSIONS

A novel method for calculating empirical orthogonal functions (EOFs) from gappy data is presented. This method uses a complex-valued extension of the Lomb-Scargle periodogram, referred to as the Lomb-Scargle Discrete Fourier Transform (LSDFT), to calculate the covariance matrix. This method can be improved with a complex-valued extension of Welch's Overlapped Segment Averaging, but a satisfactory complex-valued solution to AR1 noise reduction as per Schulz and Mudelsee (2002) remains elusive. These LSDFT EOFs can then be used to find the amplitude, phase, and variance of the primary modes of oscillation in a data set. However, a current weakness is that LSDFT EOF analysis does not explicitly yield the frequency range of each EOF, as in principal oscillation pattern (POP) analysis. The frequency dependence functions can theoretically be derived by solving the LSDFT EOF problem at each frequency instead of integrating over a bandwidth at a computational expense.

A key advantage of using LSDFT EOF analysis compared to other gappy EOF techniques like data interpolating EOF (DINEOF) analysis and pairwise-complete truncations is that all degrees of freedom contained in the data are explicitly preserved up to the integration over bandwidths. Therefore, there is no hard limit to the accuracy of LSDFT EOF analysis as it is formulated in this text. However, practical limits may make other options with hard limits on accuracy more appealing, like using NFFTs instead of the LSDFT for greater frequency resolution relative to computational cost.

Three LSDFT techniques (classic, WOSA, and AR1-corrected) are demonstrated with data from the Tropical Rainfall Measurement Mission (TRMM). All three successfully demonstrate amplitude and phase patterns over South America for the diurnal cycle,

although the AR1-corrected spectrum requires a more robust complex formulation. These spectral analyses are shown to be basically successful in identifying real-world oscillations over a bandwidth of zero to twelve days, most prominently the diurnal cycle, but also a mode consistent with cold surges. It is also proven that the LSDFT EOF method yields results that are nearly identical to complex Hilbert EOF analysis, but with the added benefit that LSDFT EOF analysis can accommodate gappy data.

Finally, the relative speeds of the programming languages R and CUDA C/C++ on a per core, per clock speed basis for this particular problem indicate that R may not be as slow for particular problems as previously thought, and only takes about 2.2 ± 0.1 times as long as CUDA in this particular situation.

REFERENCES

- Aksoy, Hafzullah. "Use of gamma distribution in hydrological analysis." *Turkish Journal of Engineering and Environmental Sciences* 24, no. 6 (2000): 419-428.
- Alvera-Azcárate, Aïda, Alexander Barth, Damien Sirjacobs, Fabian Lenartz, and Jean-Marie Beckers. "Data interpolating empirical orthogonal functions (DINEOF): A tool for geophysical data analyses." *Mediterranean Marine Science* 12, no. 3 (2011): 5-11.
- Amirataee, Babak, and Majid Montaseri. "Evaluation of L-moment and PPCC method to determine the best regional distribution of monthly rainfall data (case study: northwest of Iran)." *Journal of Urban and Environmental Engineering (JUEE)* 7, no. 2 (2013).
- Aruoba, S. Borağan, and Jesús Fernández-Villaverde. "A comparison of programming languages in economics." NBER Working Paper No. 20263 (2014), National Bureau of Economic Research.
- Beckers, Jean-Marie, and M. Rixen. "EOF calculations and data filling from incomplete oceanographic datasets." *Journal of Atmospheric and Oceanic Technology* 20, no. 12 (2003): 1839-1856.
- Bretthorst, Larry G., and Eric D. Feigelson. "Frequency estimation and generalized Lomb-Scargle periodograms." In *Statistical Challenges in Astronomy*, 309-329. New York: Springer, 2003.
- Brunet, Gilbert. "Empirical normal-mode analysis of atmospheric data." *Journal of the Atmospheric Sciences* 51, no. 7 (1994): 932-952.
- Brunet, Gilbert, and Robert Vautard. "Empirical normal modes versus empirical orthogonal functions for statistical prediction." *Journal of the Atmospheric Sciences* 53, no. 23 (1996): 3468-3489.
- Dutt, A., and V. Rokhlin. "Fast Fourier transforms for nonequispaced data." *SIAM Journal on Scientific Computing* 14, no. 6 (1993): 1368-393.
- Hanson, Lars S., and Richard Vogel. "The probability distribution of daily rainfall in the United States." In *World Environment and Water Resources Congress*, pp. 1-10. 2008.
- Hasselmann, K. "PIPs and POPs: The reduction of complex dynamical systems using principal interaction and oscillation patterns." *Journal of Geophysical Research* 93, no. 11 (1988): 015-11.
- Horel, John D. "Complex principal component analysis: Theory and examples." *Journal of Climate and Applied Meteorology* 23, no. 12 (1984): 1660-1673.

Horne, J. H., and S. L. Baliunas. "A prescription for period analysis of unevenly sampled time series." *The Astrophysical Journal* 302 (1986): 757-63.

Husak, Gregory J., Joel Michaelsen, and Chris Funk. "Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications." *International Journal of Climatology* 27, no. 7 (2007): 935-944.

Kedem, Benjamin, Long S. Chiu, and Gerald R. North. "Estimation of mean rain rate: Application to satellite observations." *Journal of Geophysical Research* 95 (1990): 1965-1972.

Keiner, Jens, Stefan Kunis, and Daniel Potts. "Using NFFT 3---A software library for various nonequispaced fast Fourier transforms." *ACM Transactions on Mathematical Software (TOMS)* 36, no. 4 (2009): 19.

Kim, Kwang Y., and Gerald R. North. "EOF analysis of surface temperature field in a stochastic climate model." *Journal of Climate* 6, no. 9 (1993): 1681-1690.

Kim, Kwang-Y., Gerald R. North, and Jianping Huang. "EOFs of one-dimensional cyclostationary time series: Computations, examples, and stochastic modeling." *Journal of the Atmospheric Sciences* 53, no. 7 (1996): 1007-1017.

Kim, Kwang-Y., and Gerald R. North. "EOFs of harmonizable cyclostationary processes." *Journal of the Atmospheric Sciences* 54, no. 19 (1997): 2416-2427.

Kim, Kwang-Y., and Qigang Wu. "A comparison study of EOF techniques: Analysis of nonstationary data with periodic statistics." *Journal of Climate* 12, no. 1 (1999): 185-199.

LeRoy, B. "Fast calculation of the Lomb-Scargle periodogram using nonequispaced fast Fourier transforms." *Astronomy & Astrophysics* 545 (2012): A50.

Lomb, Nicholas R. "Least-squares frequency analysis of unequally spaced data." *Astrophysics and Space Science* 39, no. 2 (1976): 447-462.

Lorenz, Edward N. "Empirical orthogonal functions and statistical weather prediction." Statistical Forecasting Project, MIT, Scientific Report No. 1 (1956).

Lupo, Anthony R., Joseph J. Nocera, Lance F. Bosart, Eric G. Hoffman, and David J. Knight. "South American cold surges: Types, composites, and case studies." *Monthly Weather Review* 129, no. 5 (2001): 1021-1041.

Mapes, Brian E., Thomas T. Warner, and Mei Xu. "Diurnal patterns of rainfall in northwestern South America. Part III: Diurnal gravity waves and nocturnal convection offshore." *Monthly Weather Review* 131, no. 5 (2003): 830-844.

Mathias, Adolf, Florian Grond, Ramon Guardans, Detlef Seese, Miguel Canela, Hans H. Diebner, and G. Baiocchi. "Algorithms for spectral analysis of irregularly sampled time series." *Journal of Statistical Software* 11, no. 2 (2004): 1-30.

Menabde, Merab, and Murugesu Sivapalan. "Modeling of rainfall time series and extremes using bounded random cascades and Levy-stable distributions." *Water Resources Research* 36, no. 11 (2000): 3293-3300.

Monahan, Adam H., Fredolin T. Tangang, and William W. Hsieh. "A potential problem with extended EOF analysis of standing wave fields." *Atmosphere-Ocean* 37, no. 3 (1999): 241-254.

Morandat, Floréal, Brandon Hill, Leo Osvald, and Jan Vitek. "Evaluating the design of the R language." In *ECOOP 2012-Object-Oriented Programming*, pp. 104-131. Springer Berlin Heidelberg, 2012.

Mudelsee, Manfred. "TAUEST: A computer program for estimating persistence in unevenly spaced weather/climate time series." *Computers & Geosciences* 28, no. 1 (2002): 69-72.

Nickolls, John, Ian Buck, Michael Garland, and Kevin Skadron. "Scalable parallel programming with CUDA." *Queue* 6, no. 2 (2008): 40-53.

North, Gerald R., Thomas L. Bell, Robert F. Cahalan, and Fanthune J. Moeng. "Sampling errors in the estimation of empirical orthogonal functions." *Monthly Weather Review* 110, no. 7 (1982): 699-706.

North, Gerald R. "Empirical orthogonal functions and normal modes." *Journal of the Atmospheric Sciences* 41, no. 5 (1984): 879-887.

Press, William H., and George B. Rybicki. "Fast algorithm for spectral analysis of unevenly sampled data." *The Astrophysical Journal* 338 (1989): 277-80.

R Core Team. "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria (2015), accessed 18 March 2016, <https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf>.

Robinson, P. M. "Estimation of a time series model from unequally spaced data." *Stochastic Processes and their Applications* 6, no. 1 (1977): 9-24.

Şarlak, Nermin, and A. Ünal Şorman. "Gamma autoregressive models and application on the Kızılırmak basin." *Teknik Dergi* 18, no. 3 (2007): 4219-4227.

Scargle, Jeffrey D. "Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data." *The Astrophysical Journal* 263 (1982): 835-53.

Scargle, Jeffrey D. "Studies in astronomical time series analysis. III - Fourier transforms, autocorrelation functions, and cross-correlation functions of unevenly spaced data." *The Astrophysical Journal* 343 (1989): 874-87.

Schulz, Michael, and Karl Stattegger. "Spectrum: Spectral analysis of unevenly spaced paleoclimatic time series." *Computers & Geosciences* 23, no. 9 (1997): 929-45.

Schulz, Michael, and Manfred Mudelsee. "REDFIT: Estimating red-noise spectra directly from unevenly spaced paleoclimatic time series." *Computers & Geosciences* 28, no. 3 (2002): 421-426.

Sharma, Mohita Anand, and Jai Bhagwan Singh. "Use of probability distribution in rainfall analysis." *New York Science Journal* 3, no. 9 (2010): 40-49.

Steidl, Gabriele. "A note on fast Fourier transforms for nonequispaced grids." *Advances in computational mathematics* 9, no. 3-4 (1998): 337-352.

Taylor, Marc, Martin Losch, Manfred Wenzel, and Jens Schröter. "On the sensitivity of field reconstruction and prediction using empirical orthogonal functions derived from gappy data." *Journal of Climate* 26, no. 22 (2013): 9194-9205.

Toumazou, Vincent, and Jean-Francois Cretaux. "Using a Lanczos eigensolver in the computation of empirical orthogonal functions." *Monthly Weather Review* 129, no. 5 (2001): 1243-1250.

Townsend, R. H. D. "Fast calculation of the Lomb-Scargle periodogram using graphics processing units." *The Astrophysical Journal Supplement Series* 191, no. 2 (2010): 247.

Wallace, John M., and Robert E. Dickinson. "Empirical orthogonal representation of time series in the frequency domain. Part I: Theoretical considerations." *Journal of Applied Meteorology* 11, no. 6 (1972): 887-892.

Yang, Gui-Ying, and Julia Slingo. "The diurnal cycle in the tropics." *Monthly Weather Review* 129, no. 4 (2001): 784-801.

Zechmeister, M., and M. Kürster. "The generalised Lomb-Scargle periodogram: A new formalism for the floating-mean and Keplerian periodograms." *Astronomy and Astrophysics* 496, no. 2 (2009): 577-84.

Zender, C. S., "netCDF Operator (NCO) User Guide 4.5.5-alpha07," accessed 13 February 2016, <http://nco.sf.net/nco.pdf>.